# Parsimony is hard to beat.

Kenneth Rice
Department of Biology
University of Pennsylvania
and
Tandy Warnow
Department of Computer and Information Science
University of Pennsylvania

April 1, 1997

**Abstract**

The estimation of evolutionary history from biomolecular sequences is a major intellectual project in systematic biology and many methods are used to reconstruct phylogenetic (i.e. evolutionary) trees from sequence data. These methods can be compared by studying how well each performs on sequences generated by different model trees according to some model of sequence evolution. *Statistical consistency* (whether a method can recover the topology of the model tree given infinite length sequences) and *robustness* (whether a method can recover the topology of the model tree when the assumptions of the model are violated) are two important performance issues. Not much is understood about the robustness of different methods (though see [4, 34]), but experimental and analytical performance analyses have demonstrated that maximum likelihood and all reasonable distance based methods [8] are consistent with respect to inferring trees under the general Markov model, but that maximum parsimony, a popular sequence-based method, can sometimes be positively misled, thereby inferring an incorrect topology for some model trees. Felsenstein [11] gives the original proof of inconsistency, but also see [24] for more detailed studies of when parsimony is inconsistent. Although some systematic biologists and many statisticians have rejected parsimony as a consequence of these results, others use parsimony almost exclusively and distrust distance-based methods. An important consideration in comparing the two methods is the trade-off between consistency and statistical power (the sequence length needed by the method to obtain the correct topology). In this paper we compare parsimony to two distance-based methods, a popular method called *neighbor joining*, and a new method (the *double pivot* algorithm) which approximates the $L_\infty$-nearest tree, developed and studied for use in phylogenetic reconstruction by theoretical computer science researchers. Our experimental performance analysis considers the performance of these three methods on more than 190,000 sequence data sets simulated on approximately 500 model trees. Our experimental results show that on almost all model trees, maximum parsimony delivers greater statistical power than neighbor joining, and the double-pivot algorithm exhibits the least. The experiments also indicate that parsimony is almost always more accurate than the other two methods on reasonable length sequences even under adverse conditions, such as having pairs of taxa with large evolutionary distances between them and large ratios between the highest and the lowest substitution rates on the edges.

# 1 Introduction

Inferring evolutionary trees is an important problem in biology that embeds many associated optimization problems. Optimization problems related to evolutionary tree reconstruction are usually NP-hard or conjectured to be NP-hard, so that in practice heuristics are used to reconstruct trees from data sets of more than about 20 taxa. However, the scientific objective of the physical science of phylogenetic studies is not to solve a given optimization problem, but rather to recover the order of speciation or gene duplication events as represented by the topology of the true evolutionary tree. (Locating the root of the evolutionary tree is a scientifically difficult task, so that a method is considered to have been successful if it recovers the topology of the *unrooted* tree.) This means that good or poor performance with respect to optimization problems is only important to the degree that it guarantees good or poor performance with respect to topology estimation. Consequently, experimental and analytical studies within systematic biology study the accuracy of the topology estimation of different methods, rather than the accuracy of methods with respect to associated optimization problems.

Phylogenetic reconstruction methods are compared by assuming a given model of evolution, selecting a given model tree, and generating sequences on that tree. The performance literature in phylogenetics has largely focused on three issues, *consistency*, *statistical power*, and *robustness*. A method is said to be a *consistent* method for reconstructing trees under a model of sequence evolution if for all trees in that model, the method will converge on the correct topology as the sequence length increases. *Statistical power* measures the sequence length needed by a method to obtain the correct topology with high probability from sequences generated by a model tree. *Robustness* measures the ability of the method to return the correct topology as the assumptions of the model are violated.

Robustness is possibly one of the most important aspects of performance analysis, since the model assumptions are often violated (this is especially true of the assumption that the the sites evolve identically within the tree!); yet it is the least explored experimentally or analytically. The importance of consistency is obvious, and analytical studies have shown that for many stochastic models, maximum likelihood methods and all reasonable distance-based methods are consistent, but that parsimony (a popular sequence-based method) is *not* consistent. These results have led some systematic biologists (biologists who construct trees for real data sets) and many statisticians to reject parsimony.

However, saying that a method such as parsimony is "inconsistent" does not imply that it is inconsistent on *all* model trees, but only for *some* model trees. Thus, the impact of inconsistency must be considered in the light of the question: *are the trees on which the method is inconsistent likely to occur in practice?* Many biologists say that maximum parsimony works well in practice, and that the only trees on which parsimony is clearly inconsistent are biologically unrealistic. Those who favor distance methods argue that it is dangerous to rely upon inconsistent methods, and that the best distance methods (for example, Neighbor-Joining) have sufficient statistical power to handle real data sets if they are not too large. Unfortunately, the systematic biology literature contains few analytical studies of the statistical power of different methods, so that little is known theoretically. Experimental studies have examined the case of four-taxon trees fairly thoroughly, but there has been little experimental work on trees of more than twelve or so taxa. It is unwise to base predictions of good general performance on results obtained for small trees, as a recent study [8] indicated that the statistical power of popular distance-based methods may be seriously taxed on trees with significantly distant pairs of leaves in the tree. Even so, most if not all simulation studies based upon biologically realistic trees have shown that maximum likelihood (based upon the correct model of evolution) and parsimony have greater statistical power than distance-based methods.

Our interest in comparing the statistical power of each method is due to the observation that computational requirements are easier to accommodate than sequence length requirements; hence, *a method which is computationally expensive but has great statistical power may be used reliably, while a method which is computationally efficient but has low statistical power may simply be a fast way of getting the wrong answer.* Experimental performance analysis of the statistical power of different methods is critical, because analytical results tend to yield information only about the order of magnitude of the sequence length needed for accuracy, and the sequence length so predicted can also seriously overestimate the length actually needed. These analytical results essentially guarantee good performance when the sequence length exceeds some bound, but do not imply correspondingly bad performance below the bound.

In this paper we report the results of an experimental study of three methods on approximately 190,000 sets of sequences that were stochastically generated on more than 500 different model trees. The performance of each method on each model tree was explored with respect to the sequence length needed to obtain an accurate reconstruction, so that when necessary sequences of length up to 50,000 were considered. Most of our model trees are Cavender-Farris [3] trees, and hence generate binary sequences rather than nucleotide sequences. We also included a thorough analysis of the performance of these methods on a 35-taxon *"biological"* model tree which generates nucleotide (DNA) sequences and has heterogeneous rates among the sites. This model tree is based upon a subtree for the *Lamiiadae sensu stricto* clade from a recent analysis of seed plants given in [28], and allows us to consider *robustness* as well as statistical power.

We use various topologies to explore the effects of the diameter on statistical power. On each model topology we explore a wide range of substitution probabilities to understand the effects of evolutionary distance on the performance of each method. These explorations also reveal the extent to which each method tolerates large edge mutation rates. The three methods we compare in this study include the two methods used most in practice, heuristic maximum parsimony and neighbor joining (NJ), and the new double-pivot (DP) algorithm [5] of Cohen and Farach. The double-pivot algorithm obtains a 3-approximation ratio for the $L_\infty$-nearest tree problem, and it is an improvement upon the original 3-approximation algorithm for this problem obtained by Agarwala et al in [1]. Both Neighbor Joining and the Double Pivot algorithm are distance-based methods, and both are consistent methods for reconstructing evolutionary trees under the Cavender-Farris model.

We find that despite parsimony's inconsistency, it has the greatest statistical power of the three methods we study for almost all model trees, followed by Neighbor Joining, with the Double Pivot algorithm the least powerful. That the Double Pivot algorithm should be less powerful than Neighbor Joining is perhaps surprising; the Neighbor Joining method is a simple heuristic which is not known or believed to solve exactly or approximately any known optimization problem. However, this reduction in performance for the Double Pivot algorithm is supported by analytical results reported elsewhere [2, 8], and summarized in this paper. The performance of the three methods on the biological tree in which the sites evolve heterogeneously shows that parsimony exhibits the greatest robustness of the three methods. This is perhaps not surprising, since the analytical results that guarantee consistency for distance-based methods require that the sites evolve *i.i.d*, a condition that is violated in the biological tree (and indeed in most real data sets).

Studying the conditions under which each method fails to achieve accurate topology estimations from reasonable length sequences allows us to understand the conditions under which one method might be more reliable than another. Despite predictions that parsimony requires low over-all evolutionary change in the tree we find that parsimony is not affected adversely by large interspecies distances, although the two distance-based methods *are* significantly affected. We also find that all methods are affected by high mutation rates on the edges, although parsimony's performance is more seriously degraded than the distance-based methods in those conditions. However, when analyzing large trees, the negative effects of the large diameter seem to affect the distance-based methods more than the large mutation rates affect parsimony, so that parsimony's statistical power is still superior. This suggests that parsimony may be a more reliable method for inferring large divergent trees from realistic length sequences than these distance-based methods. In fact, it could be said that our study indicates that parsimony is very hard to beat!

# 2 Preliminary material

**The Cavender-Farris model of evolution** There are many Markov models of evolution that have been proposed. For the purposes of this experimental study, we examine the performance of different methods for reconstructing trees under a simple model for binary sequence evolution, called the Cavender-Farris model [3]. Let $\{0, 1\}$ denote the two states. The root is a fixed leaf and the distribution $\pi$ at the root is uniform. For each edge $e$ of a tree $T$ we have an associated *mutation probability*, which lies strictly between 0 and 0.5. Let $p : E(T) \to (0, 0.5)$ denote the associated map. Each site (positions with the sequences) evolves down the tree identically and independently, so that $p(e)$ denotes the probability that the character state in site $i$ changes at the endpoints.

**Phylogenetic methods** A *phylogenetic tree reconstruction method* is a function $\Phi$ that associates either a tree or the statement `fail` to every collection of aligned sequences, the latter indicating that the method is unable to make such a selection for the data given. There are many different methods for inferring trees, but the most important are the distance-based methods (of which Neighbor-Joining is perhaps the most often used in practice), parsimony, and maximum likelihood.

Parsimony and maximum likelihood are *sequence-based* methods which do not first transform the sequence data into distance data. Maximum likelihood seeks the tree which is most likely to have generated the data, and hence must be based upon some model of evolution. When the correct model of evolution is assumed, the maximum likelihood method is consistent and generally requires much shorter sequences than other methods to get the correct topology. The *parsimony* cost of a tree is a measure of the number of evolutionary events indicated by the sequences labelling the nodes of the tree. If we define the Hamming distance on an edge to be the number of indices for which the sequences at the endpoints differ, then the parsimony cost of a tree in which all nodes in the tree are labelled by sequences of the same length is the sum of the Hamming distances of the edges. Given a set of aligned sequences and a tree $T$ containing these sequences as labels for the leaves, the parsimony cost of $T$ is simply the parsimony cost of $T$ given its optimal labelling. Given a fixed leaf-labelled tree, computing the parsimony cost can be achieved in polynomial time [16], but the problem of finding the best tree for a given set of sequences is NP-hard, even when the sequences are binary [7, 14]. In practice, parsimony can be solved exactly for up to about 17 sequences, but beyond such sizes heuristic searches (of unknown performance) are used to find locally optimal trees. Maximum likelihood, on the other hand, has even heavier computational requirements. It is not known to be solvable in polynomial time even on a fixed topology, and computationally it is intensive enough that generally it is not used beyond sets of about 7 or 8 sequences.

Distance-based methods are also used to reconstruct trees from biomolecular data, and have shown good good performance experimentally. There is a nice relationship between inferring model trees in the general Markov model (and hence also in the Cavender-Farris model) and distance-based reconstruction, which insures that all reasonable distance-based methods are consistent. Due to space limitations, rather than explaining the details (which can be obtained from other papers, for example [33, 9, 8]), we provide here just the essential ideas.

Given a leaf-labelled tree $T$ and edge-weighting $w$, we can define the *path distance* between leaves $i$ and $j$ to be the sum of the weights of the edges in the path between $i$ and $j$. Such distances are said to be *additive*. A *distance-method* can then be thought of as a function which assigns to a given distance matrix an additive distance matrix; for example, by returning an additive distance matrix of minimum distance to the given distance matrix under some metric on distance matrices (for example, the $L^2$ metric). It is reasonable to assume that a distance method should act as the identity on additive distance matrices, and that it should be continuous. We will call such distance methods *reasonable*.

The basic technique that allows distance-based methods to have a solid analytical foundation is the use of *corrected distance transformations.* These transformations simultaneously represent the model tree (i.e. leaf-labelled tree with information about the evolutionary process governing each edge) as an edge-weighted tree, and defines distances between sequences generated on the tree, so that the following holds: *as the sequence length increases, the matrix of observed distances converges on the additive distance defined by the edge weighted tree.* Corrected distance transformations exist for the Cavender-Farris model and for the general Markov model. The consequence of these transformations is that any *reasonable* distance method is consistent! Furthermore, the requirements for being "reasonable" are quite easy to meet, and indeed just about all the popular distance-based methods (except those that are designed to infer ultrametric trees only, such as UPGMA [32]) are reasonable, and hence consistent. Indeed, this is perhaps the main reason that distance methods are popular with statisticians. Furthermore, almost all distance methods have running times that vary between $\Theta(n^2)$ and $\Theta(n^4)$ time, and all are fast enough to analyze even large data sets, so that distance methods are appealing to some systematic biologists as well.

One way of guaranteeing consistency in a distance method is to seek an additive distance matrix of minimum distance (with respect to some metric on distance matrices) from the input distance matrix. Many metrics have been considered, but all resultant optimization problems have been shown or are assumed to be NP-hard (see [6, 10, 1] for results on such problems). One optimization problem that can be approximated well (although it is nevertheless NP-hard to solve exactly) is the $L_\infty$ nearest tree problem. Here, the $L_\infty$ metric on distance matrices is defined by $L_\infty(A, B) = \max_{ij} |A_{ij} - B_{ij}|$, and the $L_\infty$-*nearest tree* problem takes as input a distance matrix $d$ and returns an additive distance matrix $D$ minimizing $L_\infty(d, D)$. Although this problem is NP-hard, it can be 3-approximated in polynomial time [1]. The 3-approximation algorithm of Agarwala *et al.* [1] is continuous and acts as the identity on additive distance matrices, and hence provides a consistent method for inferring binary model trees for the general Markov model. The use of the 3-approximation in [1] for inferring Cavender-Farris trees was proposed by Farach and Kannan in [9] and studied with respect to the variational distance (a metric on model trees, related to but not identical to topology estimation). Experimental performance analysis of this 3-approximation algorithm and a variant called the Double-Pivot algorithm was provided by Cohen and Farach in [5]. However, Cohen and Farach were concerned with the accuracy of the approximation to the $L_\infty$-nearest tree problem and other metrics between trees, and they did not address topology estimation in their performance analysis.

The distance-based method which is used the most in practice is the **Neighbor-Joining** method of Saitou and Nei [30]. This method is a simple clustering method which finds a pair of leaves which minimize an objective criterion based upon the distance matrix, makes them siblings, thus reducing the size of the matrix, and then recurses. This method is also continuous and fixes additive matrices, and hence is also consistent. Despite its simplicity, it performs very well in practice in terms of the accuracy of its topology prediction.

The statistical power of the two distance-methods, the 3-approximation algorithm and the Neighbor-Joining Method, was studied in [8] and [2], respectively, with the following results being obtained:

**Theorem 1** *(From [8]) Let $D$ be an additive $n \times n$ distance matrix defining a binary tree $T$, $d$ be a fixed distance matrix, and let $\delta = L_\infty(d, D)$. Assume that $x$ is the minimum weight of internal edges of $T$ in the edge weighting corresponding to $D$.*
(i) *A hypothetical exact algorithm for the $L_\infty$-nearest tree is guaranteed to return the topology of $T$ from $d$ if $\delta < x/4$.*
(ii) (a) *The 3-approximation algorithm for the $L_\infty$-nearest tree is guaranteed to return the topology of $T$ from $d$ if $\delta < x/8$.* (b) *For all $n$ there exists at least one $d$ with $\delta = x/6$ for which the method can err.* (c) *If $\delta \geq x/4$, the algorithm can err for every such $d$.*
(iii) *(From Atteson [2]) The Neighbor Joining Method is guaranteed to return the topology of $T$ from $d$ if $\delta < x/2$, and there exists a $d$ for any $\delta > x/2$ for which the method can err.*

The implications of this are twofold: first, the 3-approximation algorithms and the Neighbor-Joining method are both consistent on *finite* length sequences (because the corrected distances converge on the additive distances as the sequence length increases); and, second, that the Neighbor-Joining Method can be guaranteed to be correct with high probability from *shorter* sequences than even an exact algorithm for the $L_\infty$-nearest tree. Another consequence of these results is the following:

**Theorem 2** *(From [8]) If $T$ is a Cavender-Farris model tree with edge mutation probabilities in the range $[f, g]$, and if sequences of length $k'$ are generated on this tree, where*

$$k' > \frac{c' \cdot \log n}{f^2(1-2g)^{2diam(T)}}, \tag{1}$$

*and $c'$ is some constant, then with probability $1 - o(1)$ the result of applying the Agarwala et al. algorithm to corrected distances will be the true tree. The same formula (with the constant changed) exists for the Neighbor-Joining method.*

The diameter of $T$ (which we have denoted by $diam(T)$) is the number of edges in the longest path in $T$. It is clear that for small trees, $diam(T)$ is small as well, but for large trees $diam(T)$ can be quite large. In fact, it was recently shown in [8] that $diam(T)$ grows on the order of $\sqrt{n}$ for random trees under the uniform distribution, and on the order of $\log n$ for random trees under the Yule-Harding distribution.

These findings indicate that the sequence length needed for *guaranteed* accuracy by either the Neighbor-Joining method or *even* by an exact algorithm for the $L_\infty$-nearest tree will grow extremely quickly as the number of taxa in the dataset grows. However, these results must then be evaluated experimentally, since they imply on that good performance is guaranteed in certain conditions, but do not imply bad performance when those conditions do not hold. Note that this result also suggests that experimental studies which examine only small and medium sized trees may not be predictive of performance on large trees at all, or on medium sized trees which have large diameters.

# 3 Materials and Methods

Performance analyses were run on "phylofarm" (http://phylofarm.bio.upenn.edu), an eight-CPU Pentium Pro Linux cluster devoted exclusively to phylogenetic reconstruction and performance analysis. Four basic model trees were investigated under approximately 500 edge weighting models. The smallest tree, $T_1$, is a 16-taxon complete binary tree, in which all edges have the same mutation rate. This tree has the smallest diameter of all trees on 16 taxa, and thus represents a best possible case for the distance methods whose statistical power depends upon the diameter. $T_2$ is a 35-taxon tree of the *Lamiiadae sensu stricto* from a recent analysis of seed plants [28] with edges weighted by the number of substitutions estimated under optimization of the original sequence data on the tree. The rate for each site is determined independently, based upon the analysis in [28], and hence the tree has *heterogeneous rates*; thus, this is *not* a Cavender-Farris tree. Including this tree in the analysis allows us to test *robustness* to model violations. The 35-taxon tree has a diameter of 19, which also allows us to test the effect of a large diameter on performance issues. $T_3$ is the same as $T_2$ except that we make the sites evolve *i.i.d.* $T_4$ is a caterpillar of 50 leaves; that is, it consists a path of length 48 which has 50 leaves hanging off the path. The edges incident to the first and last leaves have mutation probabilities that are five times as large as the other edges.

Edge weights for $T_2$, $T_3$, and $T_4$ were normalized on (0,1), and multiple independently seeded random data sets were generated by evolving characters down each tree. The "biological" simulations for $T_2$ (our "biological" tree) used *four-state* characters and nonidentically distributed positional substitution rates $r$ obtained by optimizing character states from the original biological data along the model tree for each of 1398 positions. These positional substitution rates were normalized to (0,1) and were repeated modulo 1398 along the randomly generated sequences. Simulations on this tree were conducted as above.

Data sets were generated using the *ecat* simulator [29]. Neighbor joining trees were calculated from multiple-events-corrected distance matrices using the *neighbor* program from PHYLIP [12]. Double pivot trees were calculated using a custom program. Heuristic parsimony trees were calculated using PAUP's [35] *hsearch* routine, with 2 to 20 heuristic search replicates, tree bisection-reconnection branch swapping, and random addition order. Searches were done with and without branch-swapping on all equally parsimonious trees. All character state optimizations and tree comparisons were calculated using PAUP [35].

A series of experiments was conducted for each of the three trees and a fourth series was conducted for the 35-taxon tree with "biological" parameters. In each series, a two-dimensional parameter space was explored for the three tree reconstruction methods by varying $p$ from a level below which all methods failed to a level above which all methods failed, and by varying the sequence length $k$ from a length below which all methods failed to a maximum length of 90,000.

## 3.1 Experimental Results

Three different model tree topologies were used to explore the range within which each method was able to recover the model topology from different length sequences. $T_1$ was the easiest tree for all the methods, and all three methods were able to achieve complete accuracy on at least 90% of the replicates from very short sequences (length at most 500) on a wide range of mutation rates. The major distinction in performance between the three methods on this tree was the level of mutation that each method could accommodate for a given sequence length. Thus, as the mutation rates increased on this topology, the sequence length parsimony required grew faster than the the sequence length either NJ or DP required. Consequently, parsimony's performance is inferior to the other methods on this tree.

We now consider $T_2$ and $T_3$, 35-taxon trees with branch lengths and topology based upon an analysis of a biological data set. $T_2$ has four-state sequences in which the sites evolve under different rates, and $T_3$ has binary sequences with *i.i.d* site evolution. Parsimony exhibits significantly better statistical power than the other two methods on both $T_2$ and $T_3$ but its advantage is more significant in $T_2$ where parsimony can achieve at least 90% accuracy from sequences of medium (5000) length for a very wide range of mutation rates, whereas DP doesn't get such levels of accuracy on any mutation rate and neighbor joining only achieves that level on a few points.

Parsimony's advantage over NJ and DP is even more pronounced on $T_4$, where parsimony is able to achieve accuracy on at least 90% of the replicates from short sequences in a very large range (.035 to .35). Neighbor-Joining never achieves 90% accuracy from short sequences, and achieves 90% accuracy on medium length (i.e. length 5000) only in a very short range (.01-.05). The performance of the Double Pivot algorithm on this tree is even worse. It achieves 90% accuracy only from long (length 10,000) sequences, and then only on one of our data points (.01). Even on very long (length 29,000) sequences, DP can only obtain this level of accuracy on a very narrow range of mutation rates (.01-.016).

In the appendix we provide the figures describing the performance of the three methods on tree $T_3$, the *i.i.d.* version of the 35 taxon biological tree. On this tree, parsimony exhibits a distinct advantage over the other two methods, but not as great an advantage as on trees $T_2$ and $T_4$. Although parsimony's performance on $T_1$ is strong, it is not as strong as the two distance-based methods there. However, $T_1$ is a very small tree, and it has a very small diameter, and hence we expect the performance on $T_3$ to be indicative of the advantage we might expect for parsimony over the two distance-based methods on larger trees.

Table 1 describes the methods in terms of the range of mutation rates they can accomodate on each tree, for four different sequence lengths ("short", "medium", "long," and "very long"). On all but $T_1$, at every sequence length parsimony can reconstruct trees from a wider range of mutation rates than NJ, and NJ can reconstruct trees from a wider range of mutation rates than DP. The comparison on $T_1$ is more subtle. All three methods show significant statistical power on $T_1$, the 16-taxon caterpillar, and each can achieve a high level of accuracy from short sequences on a wide range of mutation rates. At sequence lengths below the largest we examined, 16K, parsimony is somewhat better at detecting short edges than NJ and DP, but NJ and DP can reconstruct trees with somewhat larger edges. However, given very long sequences (16K length), parsimony is as powerful as NJ, and both outperform DP. NJ outperforms DP consistently on this tree.

## 3.2 Conclusions

On all but one tree topology (the 16 taxon complete binary tree), parsimony's statistical power was significantly greater than the other two methods, with the greatest improvement visible as the maximum mutation rate increased in the tree. The performance on the 16 taxon tree, in which parsimony does not do as well as the two distance-based methods, suggests that parsimony's statistical power *is* negatively affected by the largest edge mutation rates, more so than the two distance-based methods. However, because the statistical power of parsimony is greater than the distance-based methods' power on the larger trees, this suggests that the effects of large diameter (more precisely, of large evolutionary distance in the tree) has a greater negative effect on NJ and DP by comparison with parsimony.

Thus, we predict that in general, for trees with large evolutionary distances between some leaves, parsimony will outperform the distance-based methods, but that on smaller trees with overall smaller evolutionary distances, as the mutation rates on the edges increase, distance-based methods may do as well or better than parsimony.

We note that the performance of parsimony that we have reported is based upon heuristic searches, and because of the experimental conditions we have elected to use very few random searches through the space of trees. Improved performance for parsimony can almost certainly be obtained through more thorough searches of the tree space, as previous studies have indicated [28]. However, even when based on as few as two heuristic search replicates, parsimony seems to be able to handle higher levels of evolutionary divergence in large model trees than the other methods can.

## 4 Related work

There have been many experimental performance analysis papers in systematic biology, and most use the following methodology. Sets of sequences are obtained on some model tree, using different random seeds, so that the performance of a given method for reconstructing the topology of the model tree can be assessed. Different methods are then used to analyze the set of sequences, and their performance is evaluated with respect to the accuracy of the topology estimation. Two different types of measurements have been proposed in this regard. The first type considers the method to have been successful only if the resultant topology is identical to the topology of the model tree. The second type considers accuracy to be a measure of the topological distance between two trees, typically measuring this distance in terms of the number of edge-contractions and expansions it takes to move between the two trees (equivalently, the similarity between the two trees is the number of edge-induced bipartitions of the leaves they have in common). Despite the differences in how performance is evaluated, there are two common themes. First, the accuracy of a method is always determined by the topology of the returned tree rather than any other criterion, and second, sequences are always simulated on a model tree. Thus, even when considering the performance of distance-based methods, the method proceeds by first generating sequences on a model tree, calculating distances, and then applying the distance-based method. This methodology ensures that the results can be interpreted rigorously in terms of performance for inferring trees under some model of evolution. There are two basic questions that these papers have addressed: *consistency* and *statistical power*. Thus, studies such as [11] have shown that parsimony can be mislead by some model trees, but that all distance methods are consistent. The issue of statistical power is more complicated and is not yet well understood.

There are many experimental performance analysis papers for phylogeny estimation from which we have selected three that will enable an interesting comparison between types of studies (see [17, 18, 19, 20, 21, 22, 23, 25, 31] for some representative examples of work in this area). Of the many papers that have examined the performance of methods on four-taxon trees, we have selected [20], an insightful letter by Hillis *et al.* in *Nature,* which addresses the trade-off between statistical power and consistency. While there have been several papers examining the performance of different methods on somewhat larger data sets (up to size about twelve), few have looked at large trees. However, a recent paper by Hillis [18] examines different methods with respect to statistical power for inferring the topology of a single large (228 taxon) tree with fixed mutation rates on the edges. We also discuss a paper [5] from the literature in theoretical computer science by Cohen and Farach, which reports on an experimental performance analysis of different distance-based methods, not with respect to topology estimation, but rather with respect to their accuracy in solving certain optimization problems related to distance matrices. We include this paper because it is the only experimental study of the Double Pivot method in the literature, but we note that it uses a different methodology and seems to have a different objective than the other studies.

The letter [20] by Hillis *et al.* is concerned with the statistical power required by various methods for inferring the topology of a four-taxon tree accurately. They looked at two different model trees, one (the "easy" tree) in which every edge had the same mutation probability, and the other (the "hard" tree) which had two very "long" edges on opposite sides of the tree, and where the three remaining edges were very short. They showed using simulations that parsimony had significantly greater statistical power than the best distance-based methods on the easy tree, performing correctly on 90% of the sample data sets from 600 nucleotides, while neighbor joining using the correct distance-transformation required 5,000 nucleotides to achieve the same level of accuracy. Simulations of sequences on the hard tree showed that all the methods, except for maximum likelihood based upon the correct model of evolution, were unable to reconstruct the tree except from excessively long sequences. They concluded that with respect to statistical power, although parsimony is not consistent on all trees, there may be no reasonable alternative at hand (i.e. methods which are fast and also consistent, such as distance-based methods, seem to have very poor statistical power on those trees that parsimony cannot reconstruct).

The paper by Hillis [18] looks at the performance of different methods for inferring a model tree of 228 leaves. The model tree was based upon a tree obtained by analyzing the human mitochondrial DNA from an *African Eve* analysis, and so it has heterogeneous rates across sites, edge lengths, and topology determined by the reconstruction from the original 228-taxon data set. Hillis studied the accuracy of different methods by generating sequences of different lengths on the tree and determining the percentage of the edges that were properly reconstructed by the different methods. His study showed that the accuracy of the reconstruction obtained by parsimony was higher at each sequence length than that obtained by any other method he considered, and significantly greater than Neighbor-Joining based upon correctly computed distances.

Hillis' comment on why parsimony was able to achieve a correct topology estimation was that the amount of homoplasy in the tree was very small, indicating that he perhaps believes that trees that are both large and have high homoplasy will be hard for parsimony.

In [5], Cohen and Farach introduced the Double-Pivot algorithm, a variant of the 3-approximation algorithm for the $L_\infty$-nearest tree of Agarwala *et al.* [1], and studied its performance as compared to the original algorithm in [1] and the Neighbor-Joining Method of Saitou and Nei [30]. Their study examined a total of 60 different distance matrices, ranging in size from 6 to 20, and showed that the Double-Pivot algorithm outperformed both other methods in terms of its accuracy in solving the $L_\infty$-nearest tree and other distance-based optimization problems. Their distance matrices included some matrices from the systematic biology literature, as well as some distance matrices they generated randomly via a process in which random noise is added independently to the entries of an additive matrix.

**Commentary** There are many methodological issues which we find interesting to address via a comparison between our studies and these three methods.

The first is the distinction between studies that look at only small trees, as opposed to those (such as ours) that also examine large trees. As we showed in Theorem 2, the performance of different methods can be significantly inferior when analyzing data sets generated on trees containing pairs of sequences whose evolutionary distance is great. Our experiments also reveal this tendency for the statistical power of some methods (especially for the Double-Pivot algorithm, but also for Neighbor-Joining to a somewhat lesser degree) to decline rapidly as the maximum interleaf distance increases. In particular, we note that the sensitivity of the Double-Pivot algorithm to large evolutionary distances in the tree is not evident on the 16-taxon caterpillar, although it becomes very evident as soon as we go to the 35-taxon tree. Because the distance matrices analyzed by Cohen and Farach in [5] did not include any matrix beyond size twenty, the good performance of DP with respect to distance-based problems does not necessarily predict good performance on larger trees. Similarly, four-taxon studies, including the one by Hillis *et al.* do not necessarily provide reliable performance predictions (good or bad) on larger trees. This is a general problem with many experimental studies in systematic biology, since very few have examined larger trees. However, the main contribution of [20] is that it makes a convincing argument about the the importance of statistical power, and the importance of thinking about the trade-off between consistency and statistical power; its objective is not to imply that parsimony will be superior to NJ on large trees as well.

There are methodological differences between the Cohen and Farach [5] study and the other studies which we now discuss. The Cohen and Farach paper is not concerned with topology estimation, and in fact it is not possible to infer anything definitive about topology estimation from looking at the data in their paper. Still, it would be easy to believe that good performance with respect to various distance-based optimization problems should imply good performance with respect to topology estimation. However, if this is so, then the reported good performance of the Double Pivot algorithm over the Neighbor-Joining method with respect to optimization problems would imply good performance for DP over NJ with respect to topology estimation, contradicting our experimental results. The resolution of this apparent contradiction is that it is not clear whether superior performance with respect to the $L_\infty$-nearest tree problem implies anything about improved topology estimations. In fact, experimental studies [31, 26] about similar comparisons have shown that finding the Minimum Evolution tree may lead to worse topology estimations than using Neighbor-Joining. Although both problems are related by asymptotic results (for long enough sequences, solving such distance-based optimization problems ensures a correct topology) the question of performance on finite length sequences is more complicated and the outcome less settled. Therefore relating good performance with respect to one criterion to good performance with respect to the other is apparently *not* straightforward. Indeed, this is why other phylogenetic performance studies, including this one, have analyzed performance in terms of the accuracy of the topology estimation, rather than inferring the performance indirectly (and potentially incorrectly) through performance on distance-based optimization problems.

Working directly from distances rather than computing distances by stochastically generating sequences on model trees introduces other problems as well. Generating distance matrices that are good examples of random biological distance matrices is most reliably done by generating sequences stochastically on a model tree, and then computing distances using the usual corrected distance transformation for that model of evolution; other methods for generating "random" matrices may not produce matrices having the same properties. Indeed, it seems that modelling the *noise* introduced in the evolutionary process is a subtle statistical question, thus making it unlikely that the generation of random distance matrices through any means other than simulation of sequences will be correct. Consider, for example, the technique used by Cohen and Farach to generate random distance matrices. The random distance matrices in [5] were constructed by adding random noise independently to the entries of an additive distance matrix. However, the noise (deviation from additivity) in distances computed on sequences in a tree is *not* independent between disjoint pairs if their paths overlap in the tree. Consequently, the distance matrices used in the Cohen and Farach study are not "biological" in the sense of having the same properties as distance matrices obtained by calculating distances on sequences generated under one of these Markov processes, and hence the results reported in [5] of the Double Pivot algorithm's performance on random matrices may not be relevant to the Double Pivot's performance on random *biological* matrices.

The critiques that we have made of the Cohen and Farach study [5] arise from the distinction between methods whose objective is *clustering* versus methods whose objective is *phylogenetic reconstruction*. Clustering is a major problem which has fundamental importance in many application domains. In these domains, approximating the $L_\infty$-nearest tree may be of tremendous importance, and the experimental methodology that Cohen and Farach used may be entirely appropriate. The issues that we raise simply point out that clustering methods and performance issues may not translate immediately into this domain, in which the objective really is the reconstruction of the *topology* rather than some other optimization problem.

(For different reasons, the analytical results in [9] on the use of the 3-approximation algorithm of Agarwala *et al.* [1] for inferring Cavendar-Farris trees are also not directly comparable to our results. That study was concerned with optimizing the *Variational Distance*, a related but different objective than obtaining the topology of the Cavender-Farris model tree.)

**Recommendations for methodology** To summarize, accurate performance analysis in phylogenetic methods should (1) simulate sequences on model trees, (2) vary the model trees systematically in ways that explore the desired parameter space, being especially careful to include both large and small model tree topologies, and to vary the mutation rates, sequence lengths, and edge weight ratios. The performance criterion should be tightly coupled to topology estimation rather than to a related optimization problem, unless it is possible to establish an *a priori* relationship between solutions to one problem and solutions to the other.

# 5 When does parsimony succeed?

We note that the observations made in the two Hillis papers are congruent; parsimony has significant statistical power on those trees on which it is consistent. Because this is also in general the belief in the systematic biology community, there have been some attempts to understand the conditions in which parsimony is consistent. These experimental studies can in some sense be considered as attempts to understand the conditions under which parsimony and other methods will succeed or fail.

Two of the assumptions that have been made about the conditions under which parsimony will succeed are: Assumption 1: *Parsimony requires that the overall amount of evolutionary change in the tree per site be small.* Assumption 2: *Parsimony requires that the ratio of largest to smallest mutation probability be small.* Both our results and the two papers we discussed ([20] and [18]) address the first assumption. In [20], Hillis *et al.* address the first assumption directly by noting that parsimony does well on the "easy" tree, even though the average number of changes per site is large – 5.2. However, in [18], Hillis then seems to reverse his earlier judgement by claiming that the reason that parsimony does well on the 228 taxon tree is the fact that there are not very many changes per site (i.e. the *homoplasy* in the tree is low). However, our studies reveal that this is not a critical issue, and in fact parsimony does not seem affected by the overall homoplasy in the tree. Trees $T_2$, $T_3$, and $T_4$ when given high mutation rates both have significant homoplasy, and yet parsimony's performance does not degrade on those trees. Thus, it seems that overall evolutionary divergence is not a problem for parsimony, although it seems to be a significant problem for both Neighbor-Joining and the Double Pivot algorithms.

The reason for the second assumption is the difficult cases of four-taxon trees have been cases where the ratio between the largest and the smallest edge mutation probabilities is large. However, on large trees this again does not seem to be a problem. The ratios in our trees between the largest and the smallest mutation rates have varied from 1 for $T_1$, in which all the edges have the same rate, to 5 for $T_4$, to 36 for $T_2$ and $T_3$. Parsimony's excellent performance on $T_2$ and $T_3$ suggests that large ratios within the tree are not a problem by themselves.

The conditions under which parsimony can succeed are not clear, and the experiments (both ours and those in previous studies) indicate that parsimony is remarkably efficient on a wide range of model trees. We suspect that there are local conditions of model trees that will ensure consistency for parsimony, but we do not yet know what these conditions are. Clearly, experimental performance analysis can help us answer some of these questions.

## References


[1] R. Agarwala, V. Bafna, M. Farach, B. Narayanan, M. Paterson, and M. Thorup. On the approximability of numerical taxonomy: fitting distances by tree metrics. *Proceedings of the 7th Annual ACM-SIAM Symposium on Discrete Algorithms,* 1996.

[2] K. Atteson. The performance of the Neighbor-Joining Method of Phylogeny Reconstruction, manuscript, 1996. (Submitted to COCOON '97)

[3] Cavender, J. Taxonomy with confidence, *Mathematical Biosciences*, 40:271-280, 1978.

[4] Joseph T. Chang, Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters, submitted.

[5] J. Cohen and M. Farach, Numerical Taxonomy on Data: Experimental Results. SODA '97 and RECOMB '97.

[6] W.H.E. Day, Computational complexity of inferring phylogenies from dissimilarity matrices. *Bull. of Math. Biol.* 49(4):461-467, 1987.

[7] W.H.E. Day and D.S. Johnson, The computational complexity of inferring rooted phylogenies by parsimony. *Mathematical Biosciences*, 81:33-42, 1986.

[8] P. L. Erdős, M. A. Steel, L. A. Székely, and T. J. Warnow, Constructing big trees from short sequences. Submitted for publication (available in postscript in the home page of Székely at http://math.sc.edu/).

[9] M. Farach, and S. Kannan, Efficient algorithms for inverting evolution, *Proceedings of the ACM Symposium on the Foundations of Computer Science*, 230–236, (1996).

[10] M. Farach, S. Kannan, and T. Warnow. A robust model for finding optimal evolutionary trees. Algorithmica (1995) 13: 155-179.

[11] J. Felsenstein, Cases in which parsimony or compatibility methods will be positively misleading, *Syst. Zool.* **27** (1978), 401–410.

[12] J. Felsenstein, PHYLIP – Phylogeny Inference Package (Version 3.2), *Cladistics,* 5: 164-166, 1989.

[13] W. M. Fitch, Towards defining the course of evolution. Minimum change for specific tree topology, *Syst. Zool.* **20** (1971), 406–416.

[14] L. R. Foulds, R. L. Graham, The Steiner problem in phylogeny is NP-complete, *Adv. Appl. Math.* **3**(1982), 43–49.

[15] O. Gascuel. 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(6):961-963, 1994.

[16] J. A. Hartigan, Minimum mutation fits to a given tree, *Biometrics* **29** (1973), 53–65.

[17] D. Hillis. Approaches for assessing phylogenetic accuracy. *Syst. Biol.* 44(1):3-16, 1995.

[18] D. Hillis, Inferring complex phylogenies, *Nature* Vol **383** 12 September, 1996, 130–131.

[19] D. Hillis and J. P. Huelsenbeck. To tree the truth: Biological and numerical simulations of phylogeny. Pages 55 - 67 in Nolecular Evolution of Physiological Processes (D.M. Fambrough, ed.), Rockefeller Univ. Press, New York, 1994.

[20] D. Hillis, J. Huelsenbeck, and D. Swofford, Hobgoblin of phylogenetics? Nature, Vol. 369, 1994, pp. 363-364.

[21] J. Huelsenbeck. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor-joining. *Mol. Biol. Evol.* 12(5):843-849, 1995.

[22] J. Huelsenbeck. Performance of phylogenetic methods in simulation. *Syst. Biol.* 44(1):17-48, 1995.

[23] J. P. Huelsenbeck and D. M. Hillis, Success of Phylogenetic Methods in the Four-taxon Case, *Syst Biol.*, 42:3 247-264, 1993.

[24] J. Kim, General inconsistency conditions for maximum parsimony: effects of branch length and increasing number of taxa, *Syst. Biol.* **45**(3)(1996), 363–374.

[25] M. Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11:459-468, 1994.

[26] S. Kumar. A stepwise algorithm for finding minimum evolution trees. *Mol. Biol. Evol.* 13: 584-593. 1996.

[27] J. A. Lake, A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony, *Mol. Biol. Evol.* **4**(1987), 167–191.

[28] K. Rice, M. Donoghue, and R. Olmstead, Analyzing large data sets: *rbc*L revisited, *Systematic Biology*, to appear, 1997.

[29] K. Rice. See the web page http://www.cis.upenn.edu/∼rice/progs/ecat.

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

[31] N. Saitou and M. Imanishi. Relative efficiencies of the Fitch-Margoliash, maximum parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic reconstructions in obtaining the correct tree. *Mol. Biol. Evol.* 6:514-525, 1989.

[32] Sokal, R. R. and C. D. Michener. 1958. A statistical method for evaluating systematic relationships, Univ. of Kansas Sci. Bull. 28:1409:-1438.

[33] M. A. Steel, Recovering a tree from the leaf colourations it generates under a Markov model, *Appl. Math. Lett.*, **7(2)**, 19–24 (1994).

[34] M. A. Steel, L. A. Székely, and M. D. Hendy, Reconstructing trees when sequence sites evolve at variable rate, *J. Computational Biology* **1** (2), 153–163, 1994.

[35] D. L. Swofford, PAUP: Phylogenetic analysis using parsimony, version 3.0s. Illinois Natural History Survey, Champaign. 1992.

# 6 Tables and Figures

**Figures 1, 2 and 3** Due to space limitations we are unable to present the figures comparing the performance of these methods on each of the four trees, and so we have selected $T_3$ to present. On this tree, parsimony has a distinct advantage over the other methods, but the advantage of parsimony over the other methods is *more* pronounced on $T_2$ and $T_4$ than on $T_3$. On $T_1$, the 16-taxon complete binary tree, parsimony is actually somewhat inferior to the other two methods (though all methods perform extremely well on that tree, achieving accuracy in a large range of mutation rates for short sequences). Thus, $T_3$ presents a picture of what *may* be reasonably expected to be the minimal advantage parsimony should have over the other methods when analyzing datasets of sequences generated on other than very small trees.

Figures 1, 2, and 3 describe the experimental performance of the three methods on $T_3$. The grey scale indicates the proportion of replicates at each data point that were correctly analyzed, with white space indicating that all replicates at that data point were correctly analyzed and black space indicates that none were correctly analyzed. Early exits were taken upon prolonged success (resp.) failure when a method reconstructed 16 consecutive replicate data sets correctly at a given level of mutation or (resp.) when it failed to reconstruct four or more branch-length multiplier levels for every sequence length under analysis. The horizontal axis of each figure represents sequence length and the vertical axis represents the edge weight multiplier $p$ . Each point in the figure represents 20 replicate data sets at a given edge-weight/sequence-length point. Pixels in the extrapolated success and failure regions are colored appropriately.

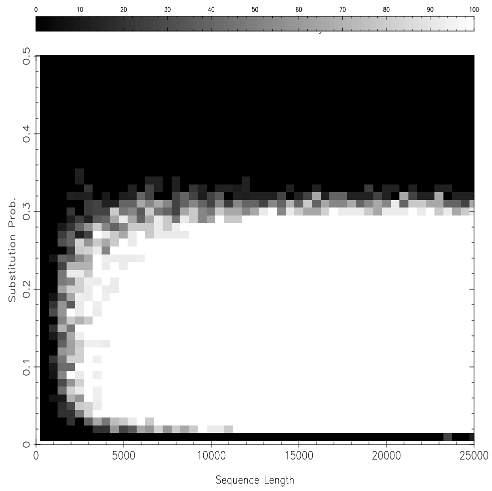


Figure 1: Heuristic Parsimony on $T_3$

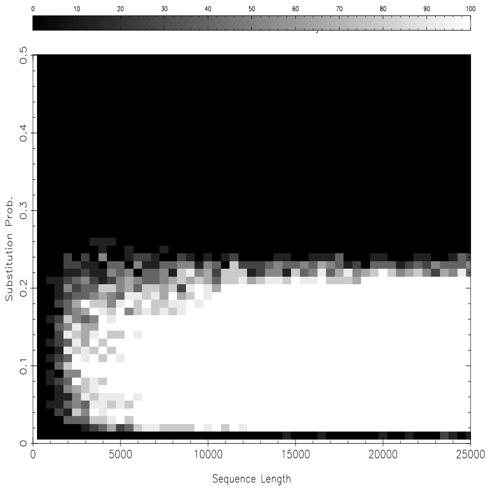


Figure 2: Neighbor Joining on $T_3$

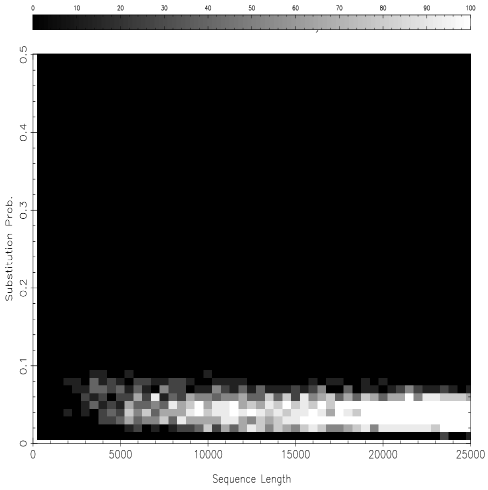


Figure 3: Double Pivot on $T_3$

| Tree | Method | short | medium | long | very long |
|---|---|---|---|---|---|
| $T_1$ | HS | .011-.181 | .011-.221 | 0.001-.221 | .001-.281 |
| | NJ | .021-.201 | .021-.281 | .001-.281 | .001-.281 |
| | DP | .021-.161 | .021-.241 | .001-.241 | .001-.261 |
| $T_2$ | HS | never | .161-.49 | .111-49 | .031-.49 |
| | NJ | never | .181-.181 | .121-.39 | .031 -.49 |
| | DP | never | never | never | .081-.191 |
| $T_3$ | HS | .151-.161 | .021-.261 | .011-.271 | .006-.296 |
| | NJ | never | .031-.166 | .011-.186 | .006-.206 |
| | DP | never | never | never | .02-.046 |
| $T_4$ | HS | .03-.35 | .01-.35 | .01-.35 | .01-.35 |
| | NJ | never | .01-.05 | .01-.05 | .01-.05 |
| | DP | never | never | .01-.01 | .01-.016 |

Table 1

Table 1 gives critical mutation substitution probability intervals between which the model tree was reconstructed for 90% of replicates. For 35-taxon and 50-taxon trees, short=1000, medium=5000, long=9000-10000, and very long=30,000. For the 16-taxon tree, short=1000, medium=6000, long=11000, and very long=16000. Rates are given as the substitution probability along the longest branch of the tree.