# Sequence-length requirements for phylogenetic methods

Bernard M.E. Moret[1], Usman Roshan[2], and Tandy Warnow[2]

[1] Department of Computer Science, University of New Mexico, Albuquerque, NM 87131
moret@cs.unm.edu
[2] Department of Computer Sciences, University of Texas, Austin, TX 78712
{usman,tandy}@cs.utexas.edu

**Abstract.** We study the sequence lengths required by neighbor-joining, greedy parsimony, and a phylogenetic reconstruction method (*DCM*$_{NJ}$+*MP*) based on disk-covering and the maximum parsimony criterion. We use extensive simulations based on random birth-death trees, with controlled deviations from ultrametricity, to collect data on the scaling of sequence-length requirements for each of the three methods as a function of the number of taxa, the rate of evolution on the tree, and the deviation from ultrametricity. Our experiments show that *DCM*$_{NJ}$+*MP* has consistently lower sequence-length requirements than the other two methods when trees of high topological accuracy are desired, although all methods require much longer sequences as the deviation from ultrametricity or the height of the tree grows. Our study has significant implications for large-scale phylogenetic reconstruction (where sequence-length requirements are a crucial factor), but also for future performance analyses in phylogenetics (since deviations from ultrametricity are proving pivotal).

## 1 Introduction

The inference of phylogenetic trees is basic to many research problems in biology and is assuming more importance as the usefulness of an evolutionary perspective is becoming more widely understood.

The inference of very large phylogenetic trees, such as might be needed for an attempt on the "Tree of Life," is a big computational challenge. Indeed, such large-scale phylogenetic analysis is currently beyond our reach: the most popular and accurate approaches are based on hard optimization problems such as maximum parsimony and maximum likelihood. Even heuristics for these optimization problems sometimes take weeks to find local optima; some datasets have been analyzed for years without a definite global optimum being found.

Polynomial-time methods would seem to offer a reasonable alternative, since they can be run to completion even on large datasets. Recent simulation studies have examined the performance of fast methods (specifically, the distance-based method neighbor-joining [21] and a simple greedy heuristic for the maximum parsimony problem) and have shown that both methods might perform rather well (with respect to topological accuracy), even on very large model trees.

These studies have focused on trees of low diameter that also obey the strong molecular clock hypothesis (so that the expected number of evolutionary events is roughly proportional to time). Our study seeks to determine whether these, and other, polynomial-time methods really do scale well with increasing number of taxa under more difficult conditions—when the model trees have large diameters and do not obey the molecular clock hypothesis.

Our paper is organized as follows. In Section 2 we review earlier studies and the issues they addressed. In Section 3, we define our terms and present the basic concepts. In Section 4 we outline our experimental design and in Section 5 we report the results of our experiments. We conclude with some remarks on the implications of our work.

## 2 Background

The *sequence-length requirement* of a method is the sequence length that this method needs in order to reconstruct the true tree topology with high probability. Earlier studies established analytical upper bounds on the sequence-length requirements of various methods, including the popular *neighbor-joining (NJ)* [21] method. These studies showed that standard methods, such as NJ, recover the true tree with high probability from sequences of lengths that are exponential in the evolutionary diameter of the true tree. Based upon these studies, we defined a parameterization of model trees in which the longest and shortest edge lengths are fixed [6, 7], so that the sequence-length requirement of a method can be expressed as a function of the number, $n$, of taxa. This parameterization led us to define *fast-converging* methods, methods that recover the true tree with high probability from sequences of lengths bounded by a polynomial in $n$ once $f$ and $g$, the minimum and maximum edge lengths, are bounded. Several fast-converging methods were subsequently developed [4, 5, 11, 24]. We and others analyzed the sequence-length requirements of standard methods, such as NJ, under the assumptions that $f$ and $g$ are fixed. These studies [1, 7] showed that NJ and many other methods can recover the true tree with high probability when given sequences of lengths bounded by a function that grows exponentially in $n$.

We then performed studies on a different parameterization of the model tree space, where we fixed the evolutionary diameter of the tree and let the number of taxa vary [17]. This parameterization, suggested to us by J. Huelsenbeck, allowed us to examine the differential performance of methods with respect to "taxon-sampling" strategies [10]. In these studies, we evaluated the performance of NJ and $DCM_{NJ}$+*MP* on simulated data, obtained from random birth-death trees with bounded deviation from ultrametricity. We found that $DCM_{NJ}$+*MP* consistently dominated NJ throughout the parameter space; we also found that the difference in performance increased as the number of taxa increased or the evolutionary rate increased.

In previous studies [15] we had studied the sequence-length requirements for NJ, Weighbor [3], Greedy MP, and $DCM_{NJ}+MP$ as a function of the numbers of taxa, with fixed height and deviation. We had found that $DCM_{NJ}+MP$ requires shorter sequences than the other four methods when trees of high topological accuracy are desired. However, we had used datasets of modest size and had not explored a very large range of evolutionary rates and deviations from ultrametricity. In this paper we remedy these limitations.

# 3 Basics

## 3.1 Simulation study

Simulation studies are the standard technique used in phylogenetic performance studies [9, 10, 14]. In a simulation study, a DNA sequence at the root of the model tree is evolved down the tree under some assumed stochastic model of evolution. This process generates a set of sequences at the leaves of the tree. The sequences are then given to phylogenetic reconstruction methods, with each method producing a tree for the set of sequences. These reconstructed trees are then compared against the model tree for topological accuracy. The process is repeated many times in order to obtain a statistically significant test of the performance of the methods under these conditions.

## 3.2 Model trees

We used random birth-death trees as model trees for our experiments. They were generated using the program r8s [22], where we specified the option to generate birth-death trees using a backward Yule (coalescent) process with waiting times taken from an exponential distribution. An edge $e$ in a tree generated by this process has length $\lambda_e$, a value that indicates the expected number of times a random site changes on the edge. Trees produced by this process are by construction ultrametric (that is, the path length from the root to each leaf is identical). Furthermore, a random site is expected to change once on any root-to-leaf path; that is, the (evolutionary) height of the tree is 1.

In our experiments we scale the edge lengths of the trees to give different heights. To scale the tree to height $h$, we multiply the length of each edge of $T$ by $h$. We also modify the edge lengths as follows in order to deviate the tree away from ultrametricity. First we pick a deviation factor, $c$; then, for each edge $e \in T$, we choose a number, $x$, uniformly at random from the interval $[-\lg(c), \lg(c)]$. We then multiply the branch length $\lambda_e$ by $e^x$. For deviation factor $c$, the expected deviation is $(c-1/c)/2lnc$, with a variance of

$\left((c^2 - \frac{1}{c^2}) - 2(c - \frac{1}{c})\right)/4\ln c$. For instance, for an expected deviation of 1.36, the standard deviation has the rather high value of 1.01.

### 3.3 DNA sequence evolution

We conducted our experiments under the Kimura 2-parameter [13] + Gamma [25] model of DNA sequence evolution, one of the standard models for studying the performance of phylogenetic reconstruction methods. In this model a site evolves down the tree under the Markov assumption and all site substitutions are partitioned into two classes: *transitions*, which substitute a purine for a purine; and *transversions*, which substitute a pyrimidine for a pyrimidine. This model has a parameter to indicate the transition/transversion ratio; in our experiments, we set it to 2 (the standard setting).

We assume that the sites have different rates of evolution drawn from a known distribution. In our experiments we use the gamma distribution with shape parameter $\alpha$ (which we set to 1, the standard setting), which is the inverse of the coefficient of variation of the substitution rate.

### 3.4 Measures of accuracy

Since the trees returned by each of the three methods are binary, we use the *Robinson-Foulds* (RF) distance [20], which is defined as follows. Every edge $e$ in a leaf-labeled tree $T$ defines a bipartition, $\pi_e$, of the leaves (induced by the deletion of $e$); the tree $T$ is uniquely encoded by the set $C(T) = \{\pi_e : e \in E(T)\}$, where $E(T)$ is the set of all internal edges of $T$. If $T$ is a model tree and $T'$ is the tree obtained by a phylogenetic reconstruction method, then the set of *False Positives* is $C(T') - C(T)$ and the set of *False Negatives* is $C(T) - C(T')$. The RF distance is then the average of the number of false positives and the number of false negatives. We plot the *RF rates* in our figures, which are obtained by dividing the RF distance by $n-3$, the number of internal edges in a binary tree on $n$ leaves. Thus the RF rate varies between 0 and 1 (or 0% and 100%). Rates below 5% are quite good, while rates above 25% are unacceptably large. We focus our attention on the sequence lengths needed to get at least 75% accuracy.

### 3.5 Phylogenetic reconstruction methods

**Neighbor-Joining** Neighbor-Joining (NJ) [21] is one of the most popular polynomial time distance-based methods [23]. The basic technique is as follows: first, a matrix of estimated distances between every pair of sequences is computed. (This step is standardized for each model of evolution, and takes $O(kn^2)$ time, where $k$ is the sequence length and $n$ the number of taxa.) The NJ method then uses a type of agglomerative clustering to construct a tree from this matrix of distances. The NJ method is provably statistically consistent under most models of evolution examined in the phylogenetics literature, and in particular under the K2P model that we use.

**Greedy Maximum Parsimony** Maximum Parsimony is an NP-hard optimization problem for phylogenetic tree reconstruction [8]. Given a set of aligned sequences (in our case, DNA sequences), the objective is to reconstruct a tree with leaves labelled by the input sequences and internal nodes labelled by other sequences so as to minimize the total "length" of the tree. This length is calculated by summing up the Hamming distances on the edges. Although the MP problem is NP-hard, many effective heuristics (most based on iterative local refinement) are offered in popular software packages. Earlier studies [19, 2] have suggested that a very simple greedy heuristic for MP might produce as accurate a reconstructed tree as the best polynomial-time methods. Our study tests this hypothesis by explicitly looking at this greedy heuristic. The greedy heuristic begins by randomly shuffling the sequences and building an optimal tree on the first four sequences. Each successive sequence is inserted into the current tree, so as to minimize the total length of the resulting, larger tree. This method takes $O(kn^2)$ time, since finding the best place to insert the next sequence takes $O(kn)$ time, where $k$ is the sequence length and $n$ is the number of taxa.

**$\mathbf{DCM_{NJ}}$+MP** The $DCM_{NJ}+MP$ method was developed by us in a series of papers [16]. In our previous simulation studies, $DCM_{NJ}+MP$ outperformed, in terms of topological accuracy, both $DCM^*_{NJ}$ (a provably fast-converging method of which $DCM_{NJ}+MP$ is a variant) and NJ. Let $d_{ij}$ represent the distance between taxa $i$ and $j$; $DCM_{NJ}+MP$ operates in two phases.

- *Phase 1:* For each choice of threshold $q \in \{d_{ij}\}$, compute a binary tree $T_q$, by using the Disk-Covering Method from [11], followed by a heuristic for refining the resultant tree into a binary tree. Let $\mathcal{T} = \{T_q : q \in \{d_{ij}\}\}$. (For details of Phase I, the reader is referred to [11].)
- *Phase 2:* Select the tree from $\mathcal{T}$ that optimizes the maximum parsimony criterion.

If we consider all $\binom{n}{2}$ values for the threshold $q$ in Phase 1, then $DCM_{NJ}+MP$ takes $O(n^6)$ time; however, if we consider only a fixed number $p$ of thresholds, it only takes $O(pn^4)$ time. In our experiments we use $p = 10$ (and $p = 5$ for 800 taxa), so that the running time of $DCM_{NJ}+MP$ is $O(n^4)$. Experiments (not shown here) indicate that choosing 10 thresholds does not reduce the performance of the method significantly (in terms of topological accuracy).

## 4 Experimental Design

We explore a broad portion of the parameter space in order to understand the factors that affect topological accuracy of the three reconstruction methods under study. In particular, we are interested in the type of datasets that might be needed in order to attempt large-scale phylogenetic analyses that span distantly related taxa, as will be needed for the "Tree of Life." Such datasets will have large heights and large numbers of taxa. Therefore we have created random model trees with heights that vary from quite small (0.05) to quite large (up to 4), and with a number of taxa ranging from 100 to 800. We have also explored the effect of deviations from the molecular clock, by using model trees that obey the molecular clock (expected deviation 1) as well as model trees with significant violations of the molecular clock (expected deviation 2.87). In order to keep the experiment tractable, we limited our analysis to datasets in which the sequence lengths are bounded by 16,000; this has the additional benefit of not exceeding by too much what we might expect to have available in a phylogenetic analysis in the coming years. In contrast the trees in [2] have smaller heights (up to a maximum of 0.5) and smaller expected deviations with less variance (their actual deviation is truncated if it exceeds a preset value), providing much milder experimental conditions than ours. The trees studied in [19] are even simpler, being ultrametric (no deviation from molecular clock) with a very small height of 0.15.

For the two distance-based methods (NJ and $DCM_{NJ}$+$MP$), we computed evolutionary distances appropriately for the K2P+gamma model with our settings for $\alpha$ and the transition/transversion ratio. Because we explore performance on datasets with high rates of evolution, we had to handle cases in which the standard distance correction cannot be applied because of saturation. In such a case, we used the technique developed in [12], called "fix-factor 1:" distances that cannot be set using the distance-correction formula are simply assigned the largest corrected distance in the matrix.

We used seqgen [18] to generate the sequences, and PAUP* 4.0 to construct Greedy MP trees as described. The software for the basic $DCM_{NJ}$ was written by D. Huson. We generated the rest of the software (a combination of C++ programs and Perl scripts) explicitly for these experiments. For job management across the cluster and public laboratory machines, we used the Condor software package.

## 5 Experimental results

We present a cross-section of our results in three major categories. All address the sequence lengths required by each reconstruction method to reach prescribed levels of topological accuracy on at least 80% of the reconstructed trees as a function of three major parameters, namely the deviation from ultrametricity, the height of the tree, and the number of taxa. We compute the sequence-length requirements to achieve a given accuracy as follows. For each method and each collection of parameter values (number of taxa, height, and deviation), we perform 40 runs to obtain 40 FN rates at each sequence length of the input dataset. We then use linear interpolation on each of the 40 curves of FN rates *vs.* sequence length to derive the sequence-length requirement at a given accuracy level. (Some curves may not reach the desired accuracy level even with sequences of length 16,000, in which case we exclude those runs.) We then average the computed sequence lengths to obtain the average sequence-length requirement at the given settings. Points not plotted in our curves indicate that less than 80% of the trees returned by the method did not have the required topological accuracy even at 16000 characters.

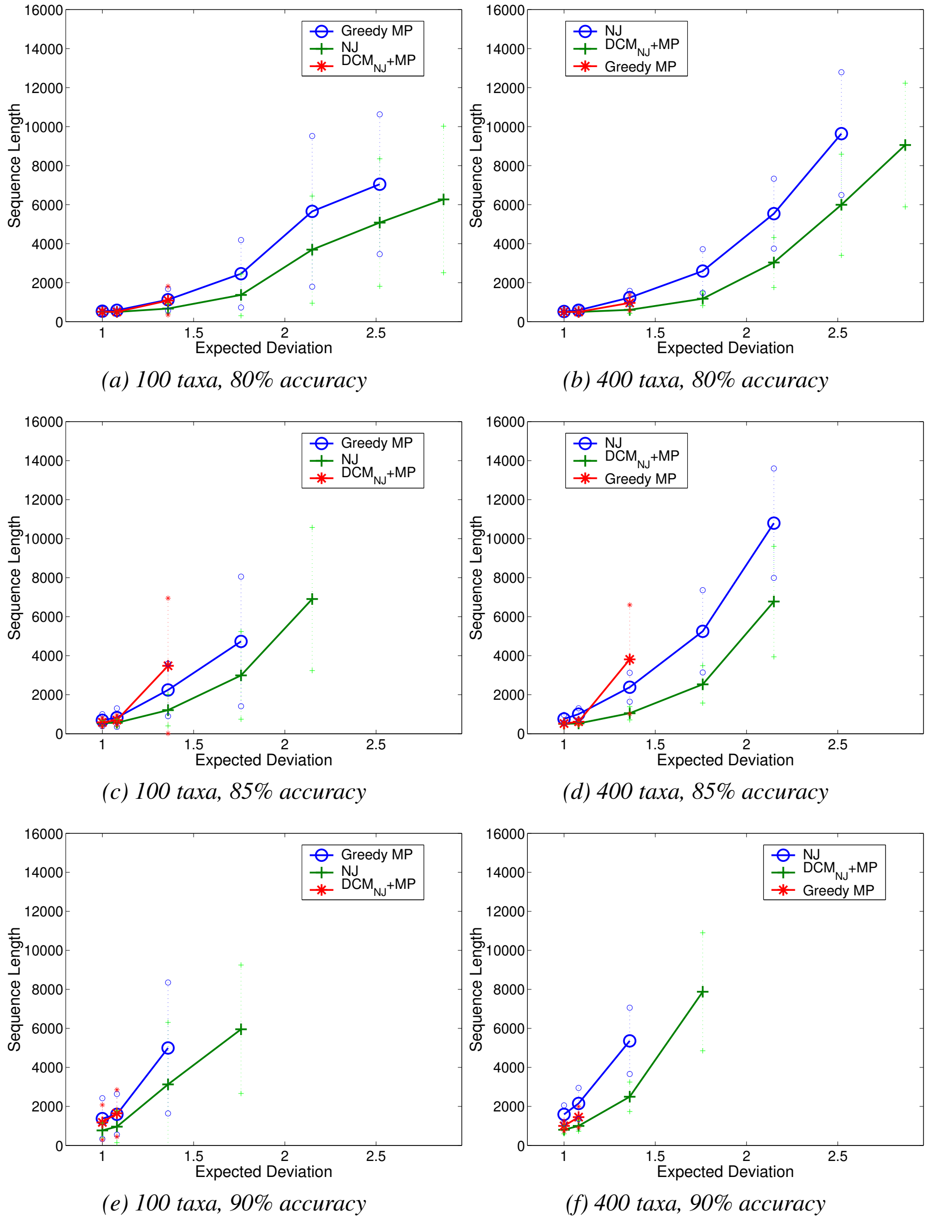
In all cases the relative requirements of the three methods follow the same ordering: $DCM_{NJ}+MP$ requires the least, followed by NJ, with greedy MP a distant third. However, even the best of the three, $DCM_{NJ}+MP$, demonstrates a very sharp, exponential rise in its requirements as the deviation from ultrametricity increases—and the worst of the three, Greedy MP, cannot be used at all with a deviation larger than 1 or with a height larger than 0.5. Similarly sharp rises are also evident when the sequence-length requirements are plotted as a function of the height of the trees. (In contrast, increases in the number of taxa cause only mild increases in the sequence length requirements, even some decreases in the case of the greedy MP method.)

### 5.1 Requirements as a function of deviation

Figure 1 shows plots of the sequence-length requirements of the three methods for various levels of accuracy as a function of the deviation. Most notable is the very sharp and fast rise in the requirements of all three methods as the deviation from ultrametricity increases. The greedy MP approach suffers so much from such increases that we could not obtain even 80% accuracy at expected deviations larger than 1.36. $DCM_{NJ}+MP$ shows the best behavior among the three methods, even though it requires unrealistically long sequences (significantly longer than 16,000 characters) at high levels of accuracy for expected deviations above 2.

### 5.2 Requirements as a function of height

Figure 2 parallels Figure 1 in all respects, but this time the variable is the height of the tree, with the expected deviation being held to a mild 1.36. The rises in




(a) *100 taxa, 80% accuracy*

(b) *400 taxa, 80% accuracy*

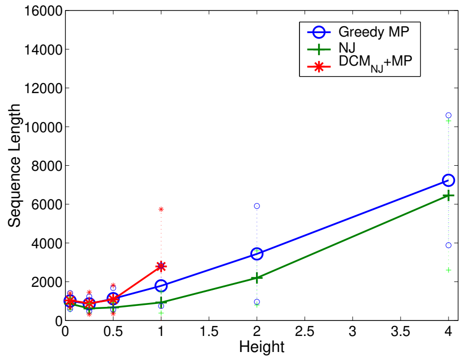
(c) *100 taxa, 85% accuracy*

(d) *400 taxa, 85% accuracy*

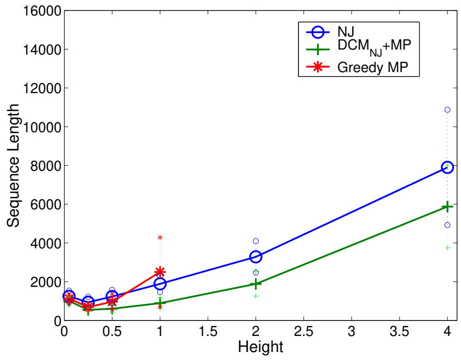
(e) *100 taxa, 90% accuracy*

(f) *400 taxa, 90% accuracy*

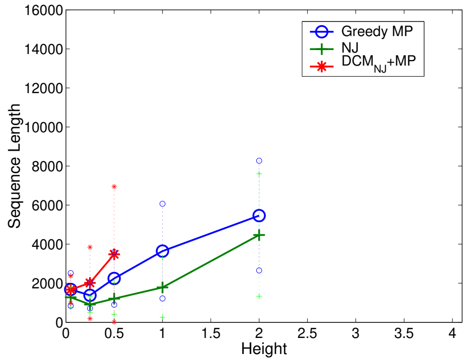
**Fig. 1.** Sequence-length requirements as a function of the deviation of the tree, for a height of 0.5, two numbers of taxa, and three levels of accuracy.



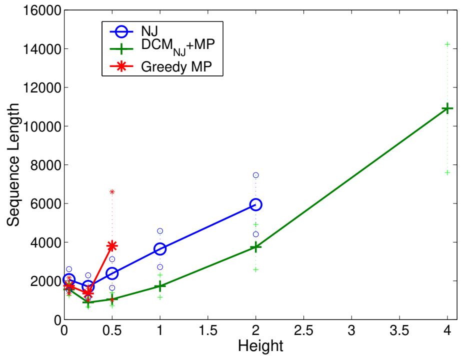

*(a) 100 taxa, 80% accuracy*



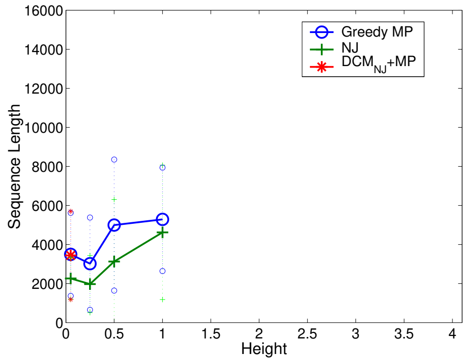

*(b) 400 taxa, 80% accuracy*



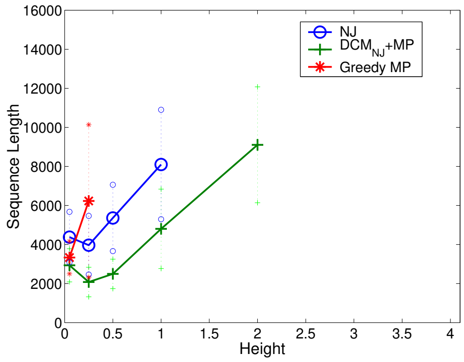

*(c) 100 taxa, 85% accuracy*




*(d) 400 taxa, 85% accuracy*

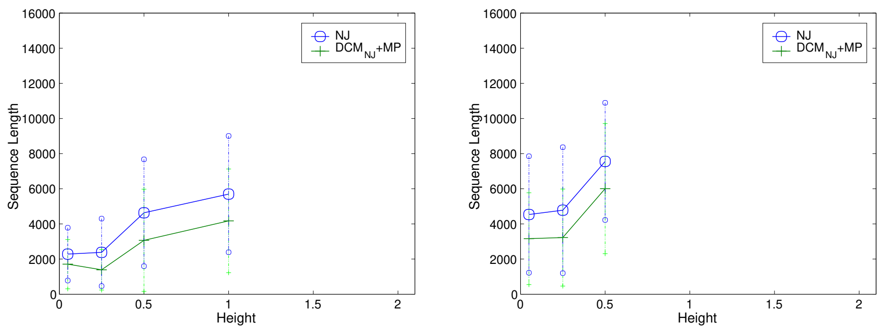



*(e) 100 taxa, 90% accuracy*




*(d) 400 taxa, 90% accuracy*

**Fig. 2.** Sequence-length requirements as a function of the height of the tree, for an expected deviation of 1.36, two numbers of taxa, and three levels of accuracy.



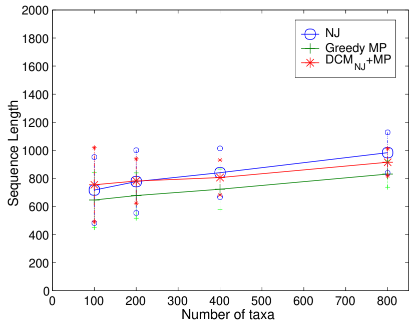

*(a) 80% accuracy*  *(b) 85% accuracy*

**Fig. 3.** Sequence-length requirements as a function of the height of the tree, for an expected deviation of 2.02, 100 taxa, and two levels of accuracy.

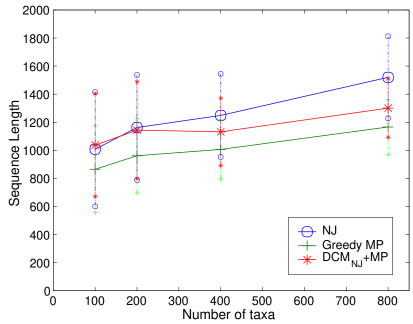
sequence-length requirements are not quite as sharp as those observed in Figure 1, but remain severe enough that accuracy levels beyond 85% are not achievable for expected deviations above 2. Once again, the best method is clearly $DCM_{NJ}+MP$ and greedy MP is clearly inadequate. Figure 3 repeats the experiment with a larger expected deviation of 2.02. At this deviation value, greedy MP cannot guarantee 80% accuracy even with strings of length 16,000, so the plots only show curves for NJ and $DCM_{NJ}+MP$, the latter again proving best. But note that the rise is now sharper, comparable to that of Figure 1. Taken together, these two figures suggest that the product of height and deviation is a very strong predictor of sequence-length requirements and that these requirements grow exponentially as a function of that product. The number of taxa does not play a significant role, although, for the larger height, we note that increasing the number of taxa decreases sequence-length requirements—a behavior that we attribute to the gradual disappearance of long tree edges under the influence of better taxon sampling.

### 5.3 Requirements as a function of the number of taxa

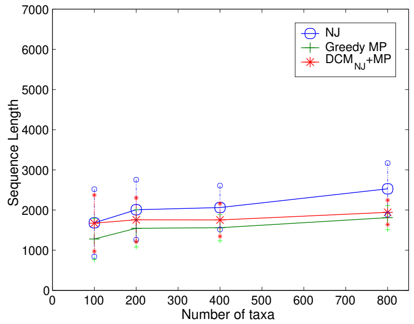
Figure 4 shows the sequence-length requirements as a function of the number of taxa over a significant range (up to 800 taxa) at four different levels of accuracy for our three methods. The tree height here is quite small and the deviation modest, so that the three methods place only modest demands on sequence length for all but the highest level of accuracy. (Note that parts (a) and (b) of the figure use a different vertical scale from that used in parts (c) and (d).) At high accuracy (90%), NJ shows distinctly worse behavior than the other two methods, with nearly twice the sequence-length requirements. Figure 5 focuses on just



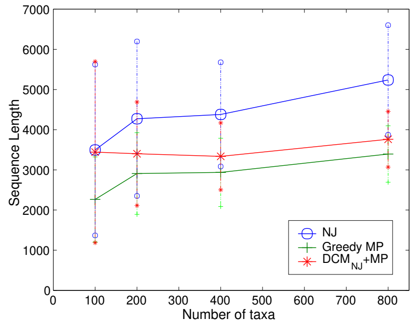

*(a) 75% accuracy*




*(b) 80% accuracy*

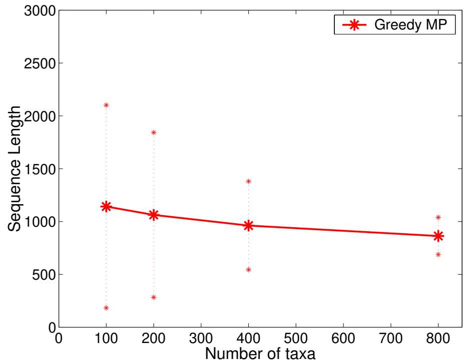



*(c) 85% accuracy*

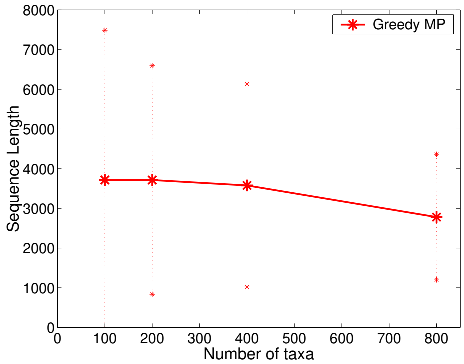



*(d) 90% accuracy*

**Fig. 4.** Sequence-length requirements as a function of the number of taxa, for an expected deviation of 1.36, a height of 0.05, and four levels of accuracy.




*(a) 80% accuracy*

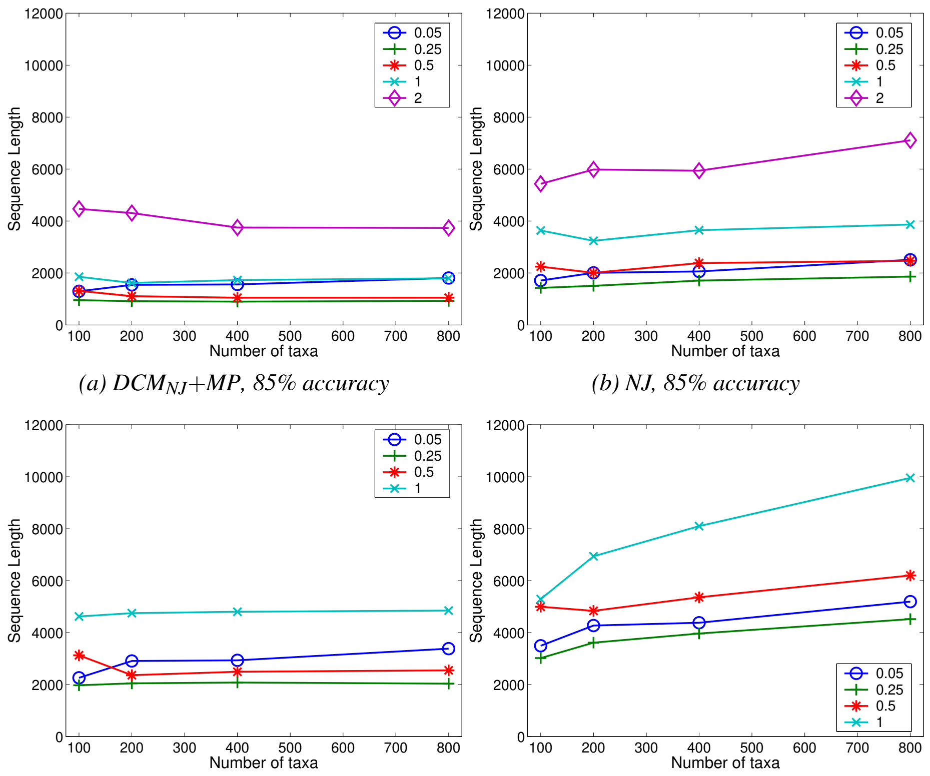



*(b) 85% accuracy*

**Fig. 5.** Sequence-length requirements for Greedy MP (60 runs) as a function of the number of taxa, for an expected deviation of 1.36, a height of 0.5, and two levels of accuracy.

the greedy MP method for trees with larger height. On such trees, initial length requirements are quite high even at 80% accuracy, but they clearly *decrease* as the number of taxa increases. This curious behavior can be explained by taxon sampling: as we increase the number of taxa for a fixed tree height and deviation, we decrease the expected (and the maximum) edge length in the model tree. Breaking up long edges in this manner avoids one of the known pitfalls for MP: its sensitivity to the simultaneous presence of both short and long edges.

Figure 6 gives a different view on our data. It shows how $DCM_{NJ}$+*MP* and NJ are affected by the number of taxa as well as by the height of the model tree. Note that the curves are not stacked in order of tree heights because Figures 2 and 3 show that the sequence-length requirements decrease from height 0.05 to 0.25 and then start to increase gradually. The juxtaposition of the graphs




*(a) $DCM_{NJ}$+MP, 85% accuracy*

*(b) NJ, 85% accuracy*

*(c) $DCM_{NJ}$+MP, 90% accuracy*

*(d) NJ, 90% accuracy*

**Fig. 6.** Sequence-length requirements (for at least 70% of trees) as a function of the number of taxa, for an expected deviation of 1.36, various heights, and two levels of accuracy, for each of $DCM_{NJ}$+*MP* and NJ.

clearly demonstrate the superiority of $DCM_{NJ}$+$MP$ over NJ: the curves for the former are both lower and flatter than those for the latter—that is, $DCM_{NJ}$+$MP$ can attain the same accuracy with significantly shorter sequences than does NJ and the growth rate in the length of the required sequences is negligible for $DCM_{NJ}$+$MP$ whereas it is clearly visible for NJ.

## 6 Conclusions and Future Work

### 6.1 Summary

We have explored the performance of three polynomial-time phylogenetic reconstruction methods, of which two (greedy maximum parsimony and neighbor-joining) are well known and widely used. In contrast to earlier studies, we have found that neighbor-joining and greedy maximum parsimony have very different performance on large trees and that both suffer substantially in the presence of significant deviations from ultrametricity. The third method, $DCM_{NJ}$+$MP$, also suffers when the model tree deviates significantly from ultrametricity, but it is significantly more robust than either NJ or greedy MP. Indeed, for large trees (containing more than 100 taxa) with large evolutionary diameters (where a site changes more than once across the tree), even small deviations from ultrametricity can be catastrophic for greedy MP and difficult for NJ.

Several inferences can be drawn from our data. First, it is clearly not the case that greedy MP and NJ are equivalent: while they both respond poorly to increasing heights and deviations from ultrametricity, greedy MP has much the worse reaction. Secondly, neither of them has good performance on large trees with high diameters—both require more characters (longer sequences) than are likely to be easily obtained. Finally, the newer method, $DCM_{NJ}$+$MP$, clearly outperforms both others and does so by a significant margin on large trees with high rates of evolution.

### 6.2 Future work

A natural question is whether a better, albeit slower, heuristic for maximum parsimony might demonstrate better behavior. Other polynomial-time methods could also be explored in this same experimental setting, and some might outperform our new method $DCM_{NJ}$+$MP$. Maximum-likelihood methods, while typically slower, should also be compared to the maximum-parsimony methods; to date, next to nothing is known about their sequence-length requirements.

Also of interest is the impact of various aspects of our experimental setup on the results. For example, a different way of deviating a model tree from the molecular clock might be more easily tolerated by these reconstruction methods—and might also be biologically more realistic. The choice of the specific model of site evolution (K2P) can also be tested, to see whether it has a significant impact on the relative performance of these (and other) methods.

## 7 Acknowledgments

This work is supported by the National Science Foundation under grants ACI 00-81404 (Moret), EIA 99-85991 (Warnow), EIA 01-13095 (Moret), EIA 01-21377 (Moret), EIA 01-21680 (Warnow), and DEB 01-20709 (Moret and Warnow), and by the David and Lucile Packard Foundation (Warnow).

## References

[1] K. Atteson. The performance of the neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25:251–278, 1999.

[2] O.R.P. Bininda-Emonds, S.G. Brady, J. Kim, and M.J. Sanderson. Scaling of accuracy in extremely large phylogenetic trees. In *Proc. 6th Pacific Symp. Biocomputing PSB 2002*, pages 547–558. World Scientific Pub., 2001.

[3] W. J. Bruno, N. Socci, and A. L. Halpern. Weighted neighbor joining: A likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.*, 17(1):189–197, 2000.

[4] M. Csűrös. Fast recovery of evolutionary trees with thousands of nodes. To appear in RECOMB 01, 2001.

[5] M. Csűrös and M. Y. Kao. Recovering evolutionary trees through harmonic greedy triplets. *Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA 99)*, pages 261–270, 1999.

[6] P. L. Erdős, M. Steel, L. Székély, and T. Warnow. A few logs suffice to build almost all trees– I. *Random Structures and Algorithms*, 14:153–184, 1997.

[7] P. L. Erdős, M. Steel, L. Székély, and T. Warnow. A few logs suffice to build almost all trees– II. *Theor. Comp. Sci.*, 221:77–118, 1999.

[8] L. R. Foulds and R. L. Graham. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics*, 3:43–49, 1982.

[9] J. Huelsenbeck. Performance of phylogenetic methods in simulation. *Syst. Biol.*, 44:17–48, 1995.

[10] J. Huelsenbeck and D. Hillis. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.*, 42:247–264, 1993.

[11] D. Huson, S. Nettles, and T. Warnow. Disk-covering, a fast-converging method for phylogenetic tree reconstruction. *Comput. Biol.*, 6:369–386, 1999.

[12] D. Huson, K. A. Smith, and T. Warnow. Correcting large distances for phylogenetic reconstruction. In *Proceedings of the 3rd Workshop on Algorithms Engineering (WAE)*, 1999. London, England.

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

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

[15] L. Nakhleh, B.M.E. Moret, U. Roshan, K. St. John, and T. Warnow. The accuracy of fast phylogenetic methods for large datasets. In *Proc. 7th Pacific Symp. Biocomputing PSB 2002*, pages 211–222. World Scientific Pub., 2002.

[16] L. Nakhleh, U. Roshan, K. St. John, J. Sun, and T. Warnow. Designing fast converging phylogenetic methods. In *Proc. 9th Int'l Conf. on Intelligent Systems for Molecular Biology (ISMB01)*, volume 17 of *Bioinformatics*, pages S190–S198. Oxford U. Press, 2001.

[17] L. Nakhleh, U. Roshan, K. St. John, J. Sun, and T. Warnow. The performance of phylogenetic methods on trees of bounded diameter. In O. Gascuel and B.M.E. Moret, editors, *Proc. 1st Int'l Workshop Algorithms in Bioinformatics (WABI'01)*, pages 214–226. Springer-Verlag, 2001.

[18] A. Rambaut and N. C. Grassly. Seq-gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comp. Appl. Biosci.*, 13:235–238, 1997.

[19] B. Rannala, J. P. Huelsenbeck, Z. Yang, and R. Nielsen. Taxon sampling and the accuracy of large phylogenies. *Syst. Biol.*, 47(4):702–719, 1998.

[20] D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

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

[22] M.J. Sanderson. `r8s` software package. Available from http://ginger.ucdavis.edu/r8s/.

[23] M.J. Sanderson, B.G. Baldwin, G. Bharathan, C.S. Campbell, D. Ferguson, J.M. Porter, C. Von Dohlen, M.F. Wojciechowski, and M.J. Donoghue. The growth of phylogenetic information and the need for a phylogenetic database. *Systematic Biology*, 42:562–568, 1993.

[24] T. Warnow, B. Moret, and K. St. John. Absolute convergence: true trees from short sequences. *Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA 01)*, pages 186–195, 2001.

[25] Z. Yang. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, 10:1396–1401, 1993.