The Accuracy of Fast Phylogenetic Methods for Large Datasets

Luay Nakhleh
Dept. of Computer Sciences
University of Texas
Austin, TX 78712

Bernard M.E. Moret
Dept. of Computer Science
University of New Mexico
Albuquerque, NM 87131

Usman Roshan
Dept. of Computer Sciences
University of Texas
Austin, TX 78712

Katherine St. John
Lehman College and
The Graduate Center
City University of New York
New York, NY 10468

Jerry Sun
Dept. of Computer Sciences
University of Texas
Austin, TX 78712

Tandy Warnow
Dept. of Computer Sciences
University of Texas
Austin, TX 78712

Whole-genome phylogenetic studies require various sources of phylogenetic signals to produce an accurate picture of the evolutionary history of a group of genomes. In particular, sequence-based reconstruction will play an important role, especially in resolving more recent events. But using sequences at the level of whole genomes means working with very large amounts of data—large numbers of sequences—as well as large phylogenetic distances, so that reconstruction methods must be both fast and robust as well as accurate. We study the accuracy, convergence rate, and speed of several fast reconstruction methods: neighbor-joining, Weighbor (a weighted version of neighbor-joining), greedy parsimony, and a new phylogenetic reconstruction method based on disk-covering and parsimony search (DCM-NJ+MP). Our study uses extensive simulations based on random birth-death trees, with controlled deviations from ultrametricity. We find that Weighbor, thanks to its sophisticated handling of probabilities, outperforms other methods for short sequences, while our new method is the best choice for sequence lengths above 100. For very large sequence lengths, all four methods have similar accuracy, so that the speed of neighbor-joining and greedy parsimony makes them the two methods of choice.

1 Introduction

Most phylogenetic reconstruction methods are designed to be used on biomolecular (i.e., DNA, RNA, or amino-acid) sequences. With the advent of gene maps for many organisms and complete sequences for smaller genomes, whole-genome approaches to phylogeny reconstruction are now being investigated. In order to produce accurate reconstructions for large collections of taxa, we will most likely need to combine both approaches—each has drawbacks not shared by the other. Because whole genomes will yield large numbers of sequences, the sequence-based algorithms will need to be very fast if they are to run within reasonable time bounds. They will also have to accommodate datasets that include very distant pairs of taxa. Many of the sequence-based reconstruction methods used by biologists (maximum likelihood, parsimony search, or quartet puzzling) are very slow and unlikely to scale up to the size of data generated in whole-genome studies. Faster methods exist (such as the popular neighbor-joining method), but most suffer from accuracy problems, especially for datasets that include distant pairs.

In this paper, we examine in detail the performance of four fast reconstruction methods, one of which we recently proposed (DCM-NJ+MP), and three others that have been used for at least a few years by biologists (neighbor-joining, Weighbor, and greedy parsimony). We ran extensive simulation studies using random birth-death trees (with deviations from ultrametricity), using about three months of computation on nearly 300 processors to conduct a thorough exploration of a rich parameter space. We used four principal parameters: model of evolution (Jukes-Cantor and Kimura 2-Parameter+Gamma), tree diameter (which indirectly captures rate of evolution), sequence length, and number of taxa. We find that Weighbor (for small sequence lengths) and our DCM-NJ+MP method (for longer sequences) are the methods of choice, although each is considerably slower than the other two methods in our study. Our data also enables us to report on the sequence-length requirements of the various methods—an important consideration, since biological sequences are of fixed length.

2 Background

Methods for inferring phylogenies are studied (both theoretically and empirically) with respect to the topological accuracy of the inferred trees. Such studies evaluate the effects of various model conditions (such as the sequence length, the rates of evolution on the tree, and the tree "shape") on the performance of the methods.

The *sequence-length requirement* of a method is the sequence length needed by the method 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 [1] method). These studies showed that standard methods, such as neighbor-joining, 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 [2,3], so that the sequence length requirement of a method can be expressed as a function of the number of taxa, n. 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 developed [4,5,6,7]. We and others analyzed the sequence length requirement of standard methods, such as neighbor-joining (NJ), under the assumptions that f and g are fixed. These studies [8,3] showed that neighbor-joining 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 recently initiated studies on a different parameterization of the model tree space, where we fix the evolutionary diameter of the tree and let the number of taxa vary [9]. This parameterization, suggested to us by J. Huelsenbeck, allows us to examine

the differential performance of methods with respect to "taxon sampling" strategies [10]. In this case, the shortest edges can be arbitrarily short, forcing the method to require unboundedly long sequences in order to recover these shortest edges. Hence, the sequence-length requirements of methods cannot be bounded. However, for a natural class of model trees, which includes random birth-death trees, we can assume $f = \Theta(1/n)$. In this case even simple polynomial-time methods converge to the true tree from sequences whose lengths are bounded by a polynomial in n. Furthermore, the degrees of the polynomials bounding the convergence rates of neighbor-joining and the fast-converging methods are identical—they differ only with respect to the leading constants. Therefore, with respect to this parameterization, there is no significant theoretical advantage between standard methods and the fast-converging methods.

In a previous study [9] we evaluated NJ and DCM-NJ+MP with respect to their performance on simulated data, obtained on random birth-death trees with bounded deviation from ultrametricity. We found that DCM-NJ+MP dominated NJ throughout the parameter space we examined and that the difference increased as the deviation from ultrametricity or the number of taxa increased.

In an unpublished study, Bruno *et al.* [11] compared Weighbor with NJ and BioNJ [12] as a function of the length of the longest edge in the true tree, using random birth-death trees of 50 taxa, deviated from the molecular clock by multiplying each edge length by a random number drawn from an exponential distribution, and using the Jukes-Cantor (JC) model of evolution. They found that Weighbor outperformed the other methods for medium to large values of the longest edge, but was inferior to them for small values—a finding we can confirm only for larger numbers of taxa. At last year's PSB, Bininda-Edmonds et al. [13] presented a study of Greedy Parsimony (which uses a single random sequence of addition and no branch swapping) in which they used very large random birth-death trees (up to 10,000 taxa), deviated from the molecular clock, and with sequences evolved under the Kimura 2-parameter (K2P) model. Unsurprisingly, they found that scaling and accuracy are at odds: the lower the accuracy level, the better the sequence length scaling.

3 Basics

3.1 Model Trees

The first step of every simulation study for phylogenetic reconstruction methods is to generate *model trees*. Sequences are then evolved down these trees, the leaf sequences are fed to the reconstruction methods under study, and the reconstructed trees compared to the original model tree.

In this paper, we use random birth-death trees with n leaves as our underlying distribution. These trees have a natural length assigned to each edge—namely, the time t between the speciation event that began that edge and the event (which could be either speciation or extinction) that ended that edge—and thus are inherently ul-

trametric. In all of our experiments we modified each edge length to deviate from this restriction, by multiplying each edge by a random number within a range $[1/c, c]$, where we set c, the *deviation factor*, to be 4.

3.2 Models of Evolution

We use two models of sequence evolution: the *Jukes-Cantor* (JC) model [14] and the the *Kimura 2-Parameter+Gamma* (K2P+Gamma) model [15]. In both models, a site evolves down the tree under the Markov assumption; in the JC model, all nucleotide substitutions (that are not the identity) are equally likely, so only one parameter is needed, whereas in the K2P model substitutions are partitioned into two classes (again other than identity): *transitions*, which substitute a purine (adenine or guanine) for a purine or a pyrimidine (cytosine or thymidine) for a pyrimidine; and *transversions*, which substitute a purine for a pyrimidine or vice versa. The K2P model has a parameter which indicates the transition/transversion ratio. We set this ratio to 2 in our experiments. Under either model, each edge of the tree is assigned a value $\lambda(e)$, the expected number of times a random site on this edge will change its nucleotide.

It is often assumed that the sites evolve identically and independently (i.i.d.) down the tree. However, we can also assume that the sites have different rates of evolution, drawn from a known distribution. One popular assumption is that the rates are drawn from a gamma distribution with shape parameter α, which is the inverse of the coefficient of variation of the substitution rate. We use $\alpha = 1$ for our experiments under K2P+Gamma.

3.3 Phylogenetic Reconstruction Methods

Neighbor Joining. Neighbor-Joining (NJ) [1] is one of the most popular distance-based methods. NJ takes a distance matrix as input and outputs a tree. For every two taxa, it determines a score, based on the distance matrix. At each step, the algorithm joins the pair with the minimum score, making a subtree whose root replaces the two chosen taxa in the matrix. The distances are recalculated to this new node, and the "joining" is repeated until only three nodes remain. These are joined to form an unrooted binary tree.

Weighted Neighbor Joining. Weighbor [16], like NJ, joins two taxa in each iteration; the pairs of taxa are chosen based on a criterion that embodies a likelihood function on the distances, which are modeled as correlated Gaussian random variables with different means and variances, computed under a probabilistic model of sequence evolution. Then, the "joining" is repeated until only three nodes remain. These are joined to form an unrooted binary tree.

DCM-NJ+MP. The DCM-NJ+MP method is a variant of a provably fast-converging method that has performed very well in previous studies [17]. In these simulation studies, DCM-NJ+MP outperformed, in terms of topological accuracy, both the provably fast converging DCM*-NJ (of which it is a variant) and NJ. We briefly describe this

method now. Let d_{ij} be the distance between taxa i and j.

- *Phase 1:* For each $q \in \{d_{ij}\}$, compute a binary tree T_q, by using the Disk-Covering Method[3], followed by a heuristic for refining the resultant tree into a binary tree. Let $\mathcal{T} = \{T_q : q \in \{d_{ij}\}\}$.
- *Phase 2:* Select the tree from $\mathcal{T}$ which optimizes the parsimony criterion.

If we consider all $\binom{n}{2}$ thresholds in Phase 1, DCM-NJ+MP takes $O(n^6)$ time, but, if we consider only a fixed number p of thresholds, it takes $O(pn^4)$ time. In our experiments, we considered only 10 thresholds, so that the running time of DCM-NJ+MP is $O(n^4)$.

Greedy Maximum Parsimony. The maximum parsimony method that we use in our study (and that was used by Bininda-Edmonds *et al.*[13]) is not, strictly speaking, a parsimony search: for the sake of speed, it uses no branch swapping at all and simply adds taxa to the tree one at a time following one random ordering of the taxa.

3.4 Measures of Accuracy

Since all the inferred trees are binary we use the *Robinson-Foulds* (RF) distance[18] which is defined as follows. Every edge e in a leaf-labeled tree T defines a bipartition π_e on the leaves (induced by the deletion of e), and hence 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 false negatives. We plot the *RF rates* in our figures, which are obtained by normalizing the RF distance by the number of internal edges in a fully resolved tree for the instance. Thus, the RF rate varies between 0 and 1 (or 0% and 100%). Rates below 5% are quite good, but rates above 20% are unacceptably large.

4 Our Experiments

In order to obtain statistically robust results, we followed the advice of[19,20] and used a number of *runs*, each composed of a number of *trials* (a trial is a single comparison), computed a mean outcome for each run, and studied the mean and standard deviation of these runs. We used 20 runs in our experiments. The standard deviation of the mean outcomes in our studies varied, depending on the number of taxa: the standard deviation of the mean on 10-taxon trees is 0.2 (which is 20%, since the possible values of the outcomes range from 0 to 1), on 25-taxon trees is 0.1 (which is 10%), whereas on 200 and 400-taxon trees the standard deviation ranged from 0.01 to 0.04 (which is between 1% and 4%). We graph the average of the mean outcomes for the runs, but omit the standard deviation from the figures.

We ran our studies on random birth-death trees generated using the r8s[21] software package. These trees have diameter 2 (height 1); in order to obtain trees with

other diameters, we multiplied the edge lengths by factors of 0.05, 0.1, 0.25 and 0.5, producing trees with diameters of 0.1, 0.2, 0.5 and 1.0, respectively. To deviate these trees from ultrametricity, we set c, the deviation factor, to 4 (see Section 3). The resulting trees have diameters at most 4 times the original diameters, and have expected diameters of 0.2, 0.4, 1.0 and 2.0. We generated such random model trees with 10, 25, 50, 100, 200, and 400 leaves, 20 trees for each combination of diameter and number of taxa.We then evolved sequences on these trees using two models of evolution, JC and K2P+Gamma (we chose $\alpha = 1$ for the shape parameter and set the transition/transversion ratio to 2). We used a fix factor[22] of 1 for distance correction. The sequence lengths that we studied are 50, 100, 250, 500, 1000 and 2000.

We used the program `Seq-Gen`[23] to generate a DNA sequence for the root and evolve it through the tree under the JC and the K2P+Gamma models of evolution. The software for DCM-NJ was written by Daniel Huson. We used PAUP* 4.0[24] for the greedy MP method, and the Weighbor 1.2 software package[16].

The experiments were run over a period of three months on about 300 different processors, all Pentiums running Linux, including the 128-processor SCOUT cluster at UT-Austin.

To generate the graphs that depict the scaling of accuracy, we linearly interpolated the sequence lengths required to achieve certain accuracy levels for fixed numbers of taxa, and then, using the interpolation, computed the sequence length, as a function of the number of taxa, that are required to achieve fixed specific accuracy levels (ones that are of interest).

5 Results and Discussion

5.1 Speed

Because we are studying methods that will scale to large datasets (large numbers of taxa and long sequences), speed is of prime importance. Table 1 shows the running time of our various methods on different instances. Note the very high speed and nearly perfect linear scaling of Greedy Parsimony. NJ is known to scale with the cube of the number of taxa; in our experiments, it scales slightly better than that. DCM-

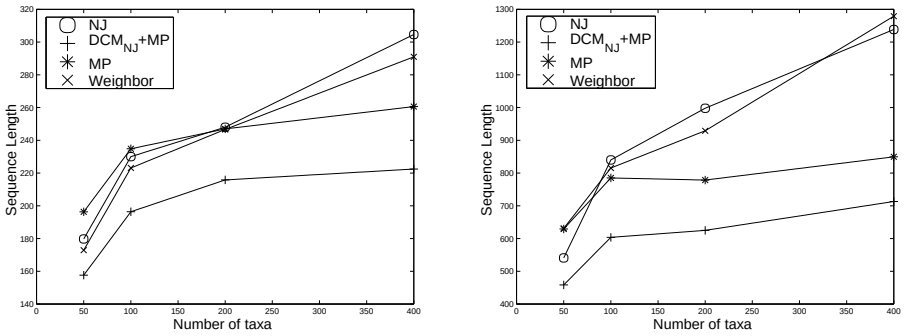
Table 1: The running times of NJ, DCM-NJ+MP, Weighbor, and Greedy MP (in seconds) for fixed sequence length (500) and diameter (0.4)

Taxa	NJ	DCM-NJ+MP	Weighbor	Greedy MP
10	0.01	1.82	0.03	0.01
25	0.02	9.12	0.37	0.02
50	0.06	21.3	3.56	0.05
100	0.37	64.25	44.93	0.10
200	2.6	470.31	352.48	0.25
400	20.13	5432.46	4077.81	0.73

NJ+MP scales exactly as NJ, but runs approximately 200 times more slowly. Finally, Weighbor scales somewhat more poorly—the figures in the table indicate scaling that is supercubic. These figures make it clear that most reasonable datasets (up to a few thousand taxa) can be processed by any of these methods—especially with the help of cluster computing, but also that very large datasets (10,000 taxa or more) will prove too costly for Weighbor and perhaps also DCM-NJ+MP (at least in their current implementations).

5.2 Sequence-Length Requirements

We can sort our experimental data in terms of accuracy and, for all datasets on which an accuracy threshold is met, count, for each fixed number of taxa, the number of datasets with a given sequence length, thereby enabling us to plot the average sequence length needed to guarantee a given maximal error rate. We show such plots for two accuracy values in Figure 1: 70% and 85%. Larger values of accuracy cannot



(a) 70% accuracy *(b) 85% accuracy*

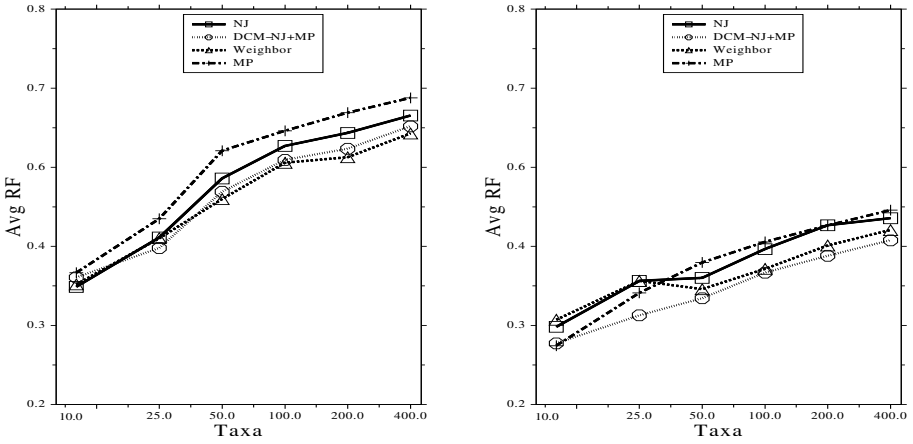
Figure 1: Sequence length requirements under the K2P+Gamma model as a function of the number of taxa

be plotted reliably, since they are rarely reached under our challenging experimental conditions. The striking feature in these plots is the difference between the two NJ-based methods (NJ and Weighbor) and the methods using parsimony (DCM-NJ+MP and Greedy Parsimony): as the number of taxa increases, the former require longer and longer sequences, growing linearly or worse, while the latter exhibit only modest growth. The divide-and-conquer strategy of DCM-NJ+MP pays off by letting its NJ component work only on significantly smaller subsets of taxa—effectively shifting the graph to the left—and completing the work with a method (parsimony) that is evidently much less demanding in terms of sequence lengths. Note that the curves are steeper for the higher accuracy requirement: as the accuracy keeps increasing, we expect to see supralinear, indeed possibly exponential, scaling.

5.3 Accuracy

We studied accuracy (in terms of the RF rate) as a function of the number of taxa, the sequence length, and the diameter of the model tree, varying one of these parameters at a time. Because accuracy varies drastically as a function of the sequence length and the number of taxa, the plots given in this section have different vertical scales.

For fixed sequence lengths and fixed diameters, we find, unsurprisingly, that the error rate of all methods increases as the number of taxa increases, although the increase is very slow (see Figures 2 and 3, but note the logarithmic scaling on the x-axis). Weighbor indeed outperforms NJ, but DCM-NJ+MP outperforms the other



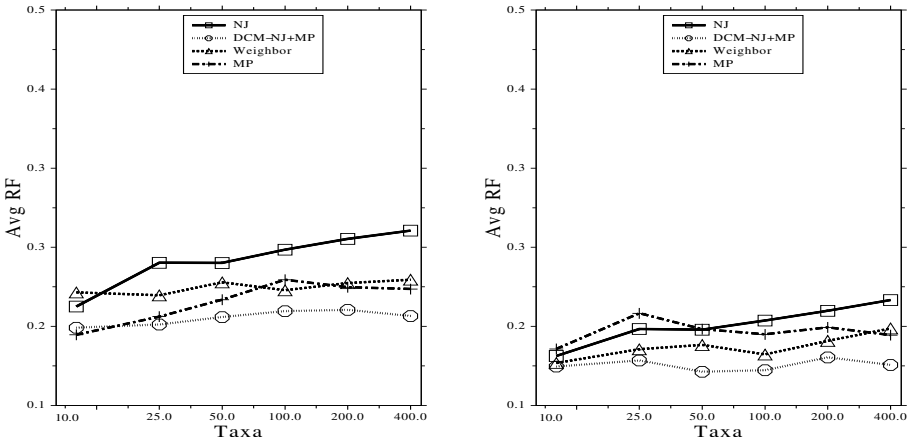
(a) sequence length 50 *(b) sequence length 100*

Figure 2: Accuracy as a function of the number of taxa under the K2P+Gamma model for expected diameter (0.4) and two sequence lengths

three methods, especially for larger trees—unless the sequences are very short, in which case Weighbor dominates.

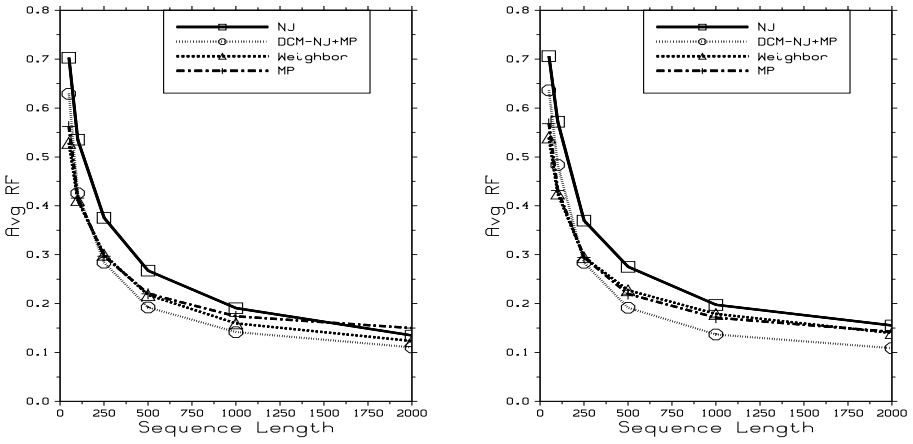
If we vary sequence length for a fixed number of taxa and fixed tree diameter, we find that the error rate decreases exponentially with the sequence length (Figure 4). From this perspective as well, DCM-NJ+MP dominates the other methods, more obviously so for larger trees. Interestingly, NJ is the worst method across almost the entire parameter space.

Finally, if we vary the diameter (which varies the rate of evolution) for a fixed number of taxa and a fixed sequence length, we find an initial increase in accuracy (due to the disappearance of zero-length edges), followed by a definite decrease (Figure 5). The decrease in accuracy is steeper with increasing diameter than what we observed with increasing number of taxa—and continually steepens. (At larger diameters—not shown, as we approach saturation, the error rate approaches unity.)



(a) sequence length 500 *(b) sequence length 1000*

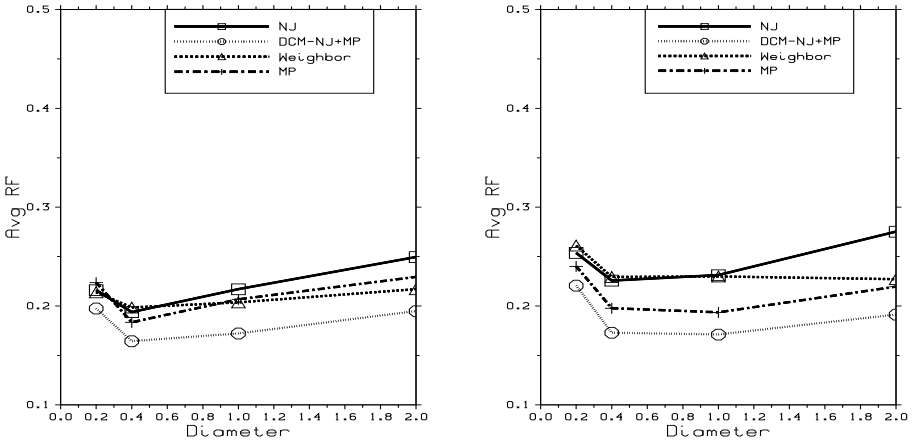
Figure 3: Accuracy as a function of the number of taxa under the K2P+Gamma model for expected diameter (2.0) and two sequence lengths



(a) 200 taxa *(a) 400 taxa*

Figure 4: Accuracy as a function of the sequence length under the K2P+Gamma model for expected diameter (2.0) and two numbers of taxa

The dominance of DCM-NJ+MP is once again evident. Comparing NJ and Weighbor, we can see that NJ is actually marginally better than Weighbor at low diameters, but Weighbor clearly dominates it at higher diameters—the two slopes are quite distinct.



(a) 100 taxa *(a) 400 taxa*

Figure 5: Accuracy as a function of the diameter under the K2P+Gamma model for fixed sequence length (500) and two numbers of taxa

5.4 The Influence of the Model of Sequence Evolution

We reported all results so far under the K2P+Gamma model only, due to space limitations. However, we explored performance under the JC (Jukes-Cantor) model as well. The relative performance of the methods we studied was the same under the JC model as under the K2P+Gamma model. However, throughout the experiments, the error rate of the methods was lower under the JC model (using the JC distance-correction formulas) than under the K2P+Gamma model of evolution (using the K2P+Gamma distance-correction formulas). This might be expected for the Weighbor method, which is optimized for the JC model, but is not as easily explained for the other methods. Figure 6 shows the error rate of NJ on trees of diameter 0.4 under the two models of evolution. NJ clearly does better under the JC model than under the K2P+Gamma model; other methods result in similar curves. Correlating the decrease in performance with specific features in the model is a challenge, but the results clearly indicate that experimentation with various models of evolution (beyond the simple JC model) is an important requirement in any study.

6 Conclusion

In earlier studies we presented the DCM-NJ+MP method and showed that it outperformed the NJ method for random trees drawn from the uniform distribution on tree topologies and branch lengths as well as for trees drawn from a more biologically realistic distribution, in which the trees are birth-death trees with a moderate deviation from ultrametricity. Here we have extended our result to include the Weighbor and

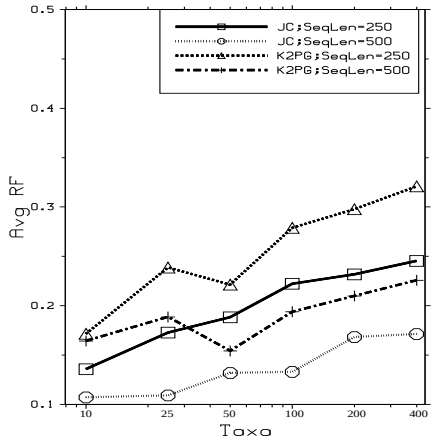


Figure 6: Accuracy of NJ as a function of the number of taxa under JC and K2P+Gamma

Greedy Parsimony methods. Our results confirm that the accuracy of the NJ method may suffer significantly on large datasets. They also indicate that Greedy Parsimony, while very fast, has mediocre to poor accuracy, while Weighbor and DCM-NJ+MP consistently return good trees, with Weighbor doing better on shorter sequences and DCM-NJ+MP doing better on longer sequences. Among interesting questions that arise are: (i) is there a way to conduct a partial parsimony search that scales no worse than quadratically (and might outperform DCM-NJ+MP)? (ii) would a DCM-Weighbor+MP prove a worthwhile tradeoff? (iii) can we make quantitative statements about the accuracy achievable by any method (not just one of those under study) as a function of some of the model parameters?

7 Acknowledgments

This work was supported in part by the National Science Foundation with grants EIA 99-85991 to T. Warnow and ACI 00-81404 to B.M.E. Moret and a POWRE award to K. St. John; by the Texas Institute for Computational and Applied Mathematics and the Center for Computational Biology at UT-Austin (K. St. John); and by the David and Lucile Packard Foundation (T. Warnow).

References

1. N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.
2. P. L. Erdos, 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.
3. P. L. Erdos, M. Steel, L. Székély, and T. Warnow. A few logs suffice to build almost all trees–

II. *Theor. Comp. Sci.*, 221:77–118, 1999.
4. M. Csűrös. Fast recovery of evolutionary trees with thousands of nodes. RECOMB 01, 2001.
5. M. Csűrös and M. Y. Kao. Recovering evolutionary trees through harmonic greedy triplets. *Proc. 10th ACM-SIAM Symp. on Discrete Algorithms (SODA 99)*, pages 261–270, 1999.
6. D. Huson, S. Nettles, and T. Warnow. Disk-covering, a fast-converging method for phylogenetic tree reconstruction. *Comput. Biol.*, 6:369–386, 1999.
7. T. Warnow, B. Moret, and K. St. John. Absolute convergence: true trees from short sequences. *Proc. 12th ACM-SIAM Symp. on Discrete Algorithms (SODA 01)*, pages 186–195, 2001.
8. K. Atteson. The performance of the neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25:251–278, 1999.
9. L. Nakhleh, U. Roshan, K. St. John, J. Sun, and T. Warnow. The performance of phylogenetic methods on trees of bounded diameter. In *Proc. 1st Workshop on Algorithms in Bioinformatics (WABI 01)*, pages 214–226, Aarhus (2001). in *LNCS* 2149.
10. J. Huelsenbeck and D. Hillis. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.*, 42:247–264, 1993.
11. http://www.t10.lanl.gov/billb/weighbor/performance.html.
12. O. Gascuel. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.*, 14:685–695, 1997.
13. O. Bininda-Emonds, S. Brady, J. Kim, and M. Sanderson. Scaling of accuracy in extremely large phylogenetic trees. In *Proc. 6th Pacific Symp. on Biocomputing (PSB01)*, pages 547–557. World Scientific, 2001.
14. T. Jukes and C. Cantor. Evolution of protein molecules. In H.N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, NY, 1969.
15. 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.
16. 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.
17. 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 Mol. Biol. (ISMB 01)*, 2001. In *Bioinformatics* 17:S190–S198.
18. D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.
19. B. M. E. Moret and H. D. Shapiro. Algorithms and experiments: the new (and the old) methodology. *J. Univ. Comput. Sci.*, 7(5):434–446, 2001.
20. C. McGeoch. Analyzing algorithms by simulation: variance reduction techniques and simulation speedups. *ACM Comp. Surveys*, 24:195–212, 1992.
21. M. Sanderson. `r8s` software package. Available from http://loco.ucdavis.edu/r8s/r8s.html.
22. D. Huson, K. A. Smith, and T. Warnow. Correcting large distances for phylogenetic reconstruction. In *Proc. 3rd Workshop on Algorithms Engineering (WAE 99)*, 1999. London, England.
23. 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.
24. D. L. Swofford. PAUP*: Phylogenetic analysis using parsimony (and other methods), 1996. Sinauer Associates, Underland, Massachusetts, Version 4.0.