

The Performance of Phylogenetic Methods on Trees of Bounded Diameter

Luay Nakhleh¹, Usman Roshan¹, Katherine St. John^{1,2}, Jerry Sun¹, and Tandy Warnow^{1,3}

¹ Department of Computer Sciences, University of Texas, Austin, TX 78712;
{nakhleh, usman, jsun, stjoh, tandyt}@cs.utexas.edu

² Department of Math & Computer Science, Lehman College & the Graduate Center, City University of New York

Supported in part by NSF Award 99-73874 and the Texas Institute for Computational and Applied Mathematics.

³ Supported by the David and Lucile Packard Foundation.

Abstract. We study the convergence rate of neighbor joining and several new phylogenetic reconstruction methods on families of trees of bounded diameter. Our study presents theoretically obtained convergence rates, as well as an empirical study based upon simulation of evolution down random birth death trees and “biological” model trees. We find that the new phylogenetic methods offer an advantage over the NJ method, except for low rates of evolution, where they have comparable performance. The improvement in performance of the new methods over NJ increase with the number of taxa and the rate of evolution.

1 Introduction

Phylogenetic trees (that is, evolutionary trees) form an important part of much biological research. There are many algorithms for inferring phylogenetic trees. The majority of these methods are designed to be used on biomolecular (i.e. DNA, RNA, or amino-acid) sequences. Methods for inferring phylogenies from biomolecular sequence data 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 various methods.

The sequence length requirement of a method is the sequence length needed by the method in order to obtain (with high probability) the true tree topology. Earlier studies established analytical upper bounds on the sequence length requirements of various methods (including the popular Neighbor Joining [14]) 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, in [6, 7] we defined a parameterization of model trees in which the longest

and shortest edge lengths are fixed, so that the sequence length requirement of a method can be expressed as a function of the number n of taxa. This parameterization lead to the definition of “fast converging” methods, which are methods that recover the true tree from sequences of lengths bounded by a polynomial in n once f , the minimum edge length, and g , the maximum edge length, are bounded. Several fast converging methods were developed [5, 4, 9, 16]. 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 [7, 1] showed that NJ and many other methods can be proven to be “exponentially converging”, that is they recover the true tree with high probability from sequences of lengths bounded by a function that grows exponentially in n , but so far none of these standard methods are known to be “fast converging”.

In this paper we consider a different parameterization of model tree space, where we fix the evolutionary diameter of the tree, and let the number of taxa vary. This parameterization, suggested by John Huelsenbeck [personal communication], allows us to examine the differential performance of methods with respect to “taxon sampling” strategies [8]. 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 all methods cannot be bounded. However, for a natural class of model trees, it can be assumed that $f = \Theta(1/n)$ (for example, random birth death trees fall into this class), and in this case we can show that even very 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 NJ 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. We then evaluate two methods, neighbor joining and DCM-NJ+MP (a method introduced in [11]) with respect to their performance on simulated data, obtained on random birth death trees with bounded deviation from ultrametricity. We find that DCM-NJ+MP obtains an advantage over neighbor joining throughout most of the parameter space we examine, and is never worse. That advantage increases as the deviation from ultrametricity increases, or as the number of taxa increases.

The rest of the paper is organized as follows. In Section 2, we present the basic definitions, models of evolution, methods, and terms, upon which the rest of the paper is based. In Section 3, we present the theory behind convergence rates bounds for both neighbor joining and “fast converging” methods. We derive bounds on the convergence rates of various methods for trees in which the

evolutionary diameter (but not the shortest edge lengths) is fixed. We then derive bounds on the convergence rates of these methods for random trees drawn from the distribution on birth-death trees described above. In Section 4, we describe our experimental study comparing the performance of neighbor joining and DCM-NJ+MP. In Section 5, we conclude with a discussion and open problems.

2 Basics

In this section, we present the basic definitions, models of evolution, methods, and terms, upon which the rest of the paper is based.

2.1 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, and these sequences are used, by the methods in question, to estimate the model tree. The accuracy of the method is determined by how well the method reproduces the model tree. Model trees are often taken from some underlying distribution on all rooted trees with n leaves. Some possible distributions include the uniform (all trees on n leaves are given equal weight) and the Yule distribution (trees are weighted by the likeliness of a speciation event occurring on external edges).

In this paper, we use random birth-death trees with n leaves as our underlying distribution. To generate these trees, we view speciation and extinction events occurring over a continuous interval. During a short time interval, δt , a species can split into two with probability $b(t)\delta t$, and a species can become extinct with probability $d(t)\delta t$. The values of $b(t)$ and $d(t)$ depend on how much time has passed in the model. To generate a tree with n taxa, we begin this process with a single node and continue until we have a tree with n taxa (with some non-zero probability some processes will not produce a tree of the desired size since all nodes could go "extinct" before n species are generated. If this happens, repeat the process, until a tree of the desired size is generated.) Under this distribution, trees have a natural length assigned to each edge— that is 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.

Birth-death trees are inherently ultrametric, that is, the branch lengths are proportional to time. In some of our experiments we modified each edge length so as to deviate from this assumption that sites evolve under the strong molecular clock. To do this we multiplied each edge by a random number within a range $[1/c, c]$, where we set c to be some small constant. We call this constant the "scaling factor."

2.2 Models of Evolution

Under the *Kimura 2-Parameter* (K2P) model, each site evolves down the tree under the Markov assumption, but there are two different types of nucleotide substitutions: transitions and transversions. The probability of a given nucleotide substitution depends on the edge and upon the type of substitution. A K2P tree is defined by the triplet $(T, \{\lambda(e)\}, ts/tv)$, where ts/tv is the transition/transversion ratio. (In our experiments, we fix this ratio to 2, one of the standard settings).

This model describes how a single site (that is, a position within the sequence at the root) evolves down the tree, and it is assumed that the sites evolve identically and independently. However, we can also assume that the sites have different rates of evolution, and that these rates are 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. With these assumptions, we can specify a K2P+Gamma tree just by the pair $(T, \{\lambda(e)\})$.

2.3 Statistical Performance Issues

Under the assumption of a K2P+Gamma evolutionary process, if the transition/transversion ratio and shape parameter are known, it is possible to define pairwise distances between taxa so that distance-based methods (such as neighbor joining) are statistically consistent [10]. (A phylogenetic reconstruction method M is *statistically consistent* under a model of evolution if for every tree in that model the probability that the method reconstructs the tree tends to 1 as the sequence length increases.) Real data are of limited length. Therefore, the length k of the sequences affects the performance of the method M significantly. The *convergence rate* of a method M is the rate at which it converges to 100% accuracy as a function of the sequence length.

2.4 Phylogenetic Reconstruction Methods

We briefly discuss the two phylogenetic methods we use in our empirical studies: neighbor joining and DCM-NJ+MP. Both methods have polynomial running time.

Neighbor Joining: Neighbor joining [14] is one of the most popular distance based methods. Neighbor takes a distance matrix as input and output 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.

DCM-NJ+MP: DCM-NJ+MP is a variant of methods proven to be fast converging that has performed very well in previous studies (see [11]). DCM-NJ+MP outperforms, in terms of topological accuracy, the methods *DCM**-NJ (of which it is a variant) and neighbor joining in these simulation studies.

The method works as follows: 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 from [7], followed by a heuristic for refining the resultant tree into a binary tree. Let $\mathcal{T} = \{T_q : q \in \{d_{ij}\}\}$. (Readers interested in more details of how Phase I is handled should see [7].)
- *Phase 2:* Select the tree from \mathcal{T} which optimizes the parsimony criterion.

DCM-NJ+MP is not statistically consistent, even under the simplest models, since the maximum parsimony criterion can select the wrong tree with probability going to 1 as the sequence length increases. However, in our earlier experimental studies, we found that DCM-NJ+MP had very good performance, returning more accurate trees than a very similar method (differing only in how Phase 2 is implemented) which is statistically consistent and “fast converging”. For this reason, we have selected DCM-NJ+MP to compare against neighbor joining in our simulation studies.

2.5 Measures of accuracy

There are many ways of measuring error between trees. Since our inferred trees are not always binary, we use the False Negative (FN) Rate to score trees. We now define this. Each edge in a tree induces a bipartition on the set of leaves of the tree. The FN error rate is the proportion of bipartitions in the true (i.e. model) tree that are missing in the inferred tree. Since the model tree is binary, when the FN error rate is 0, the inferred tree is equal to the model tree.

3 Theoretical Results on Convergence Rates

The largest and smallest edge-lengths clearly affect the sequence length needed by any method. In [16], we examined the convergence rate issue by fixing arbitrarily the largest and smallest “edge-lengths” (the length of an edge e is defined to be $\lambda(e)$, the expected number of times a random site will change its nucleotide on e). Once these bounds are fixed, we can consider the sequence

length a method needs in order to recover the tree topology exactly with high probability. This sequence length “requirement” clearly grows with the number of leaves in the tree.

In [1], the sequence length requirement for the neighbor joining method under the Cavender-Farris model was bounded from above, and extended to the General Markov model in [6]. We state the result here:

Theorem 1. *Let (T, M) be a model tree in the GM model. Let $\lambda(e) = -\log |\det(M_e)|$, and set $\lambda_{ij} = \sum_{e \in P_{ij}} \lambda(e)$. Assume that f is fixed with $0 < f \leq \lambda(e)$ for all edges $e \in T$. Let $\varepsilon > 0$ be given. Then, there are constants C and C' (that do not depend upon f) such that, for*

$$k = \frac{C}{f^2} \log n e^{C'(\max \lambda_{ij})}$$

then with probability at least $1 - \delta$, $NJ(S) = T$, where S is a set of sequences of length k generated on T . The same sequence length requirement applies to the Q^ method of [2].*

3.1 Fixed-parameter analyses of the convergence rate

Analysis when both f and g are fixed In [16, 9] we analyzed the convergence rate of NJ when both f and g are fixed (recall that f is the smallest edge length, and g is the largest edge length). In this setting, by Theorem 1 and because $\max \lambda_{ij} = O(gn)$, we see that NJ recovers the true tree with probability $1 - \varepsilon$ from sequences that grow exponentially in n . An average case analysis of tree topologies under various distributions shows that $\max \lambda_{ij} = \Theta(g\sqrt{n})$ for the uniform distribution and $\Theta(g \log n)$ for the Yule-Harding distribution. Hence, NJ has an average case convergence rate which is polynomial in n under the Yule-Harding distribution, but not under the uniform distribution.

By definition, “fast-converging” methods are required to converge to the true tree from polynomial length sequences, when f and g are fixed. The convergence rates of fast-converging methods have a somewhat different form. We show the analysis for the DCM^* -NJ method (see [16]):

Theorem 2. *Let (T, M) be a model tree in the GM model. Let $\lambda(e) = -\log |\det(M_e)|$, and set $\lambda_{ij} = \sum_{e \in P_{ij}} \lambda(e)$. Assume that f is fixed with $0 < f \leq \lambda(e)$ for all edges $e \in T$. Let $\varepsilon > 0$ be given. Then, there are constants C and C' (that do not depend upon f) such that, for*

$$k = \frac{C}{f^2} \log n e^{C'(\text{width}(T))}$$

then with probability at least $1 - \delta$, $DCM^\text{-}NJ(S) = T$, where S is a set of sequences of length k generated on T , and $\text{width}(T)$ is a topologically defined function which is bounded from above by $\max \lambda_{ij}$ and is also $O(g \log n)$.*

Consequently, fast-converging methods recover the true tree from polynomial length sequences when both f and g are fixed.

Analysis when $\max \lambda_{ij}$ is fixed. Suppose now that we fix $\max \lambda_{ij}$ but not f . In this case, neither NJ nor the “fast-converging” methods will recover the true tree from sequences that grow polynomially in n , because as $f \rightarrow 0$, the sequence length requirement increases without bound. However, for “random” birth-death trees, the expected minimum edge length is $\Theta(1/n)$. Hence, suppose that in addition to fixing $\max \lambda_{ij}$ we also require that $f = \Theta(1/n)$. In this case, an application of Theorem 1 and Theorem 2 shows that NJ and the “fast-converging” methods all recover the true tree with high probability from $O(n^2 \log n)$ length sequences. The theoretically obtained convergence rates differ only in the leading constant, which in NJ’s case depends exponentially on $\max \lambda_{ij}$, while in the case of DCM*-NJ’s this rate depends exponentially on $width(T)$. Thus, the performance advantage of a fast-converging method – from a theoretical perspective – depends upon the difference between these two values. We know that $width(T) \leq \max \lambda_{ij}$ for all trees. Furthermore, the two values are equal essentially only when the strong molecular clock assumption holds. Note also that when the tree has low evolutionary diameter (i.e. when $\max \lambda_{ij}$ is small), then the predicted performance of these methods suggests that they will be approximately identical. Only for large evolutionary diameters should we obtain a performance advantage by using the fast-converging methods instead of NJ.

In the next section we tested the empirical performance of these methods.

4 Our Performance Studies

4.1 Experimental Platform

Machines: The experiments were run on the SCOUT cluster, which contains approximately 280 different processors running the Debian Linux operating system. We also had nighttime use of approximately 150 Pentium III processors located in public undergraduate laboratories.

Software: We used Sanderson’s `r8s` package for generating birth-death trees [13]. We used the program `Seq-Gen` [12] to randomly generate a DNA sequence for the root and evolve it through the tree under K2P+Gamma model of evolution. We calculated evolutionary distances appropriately for the model (see [10]).

The software for DCM-NJ was written by Daniel Huson. To calculate the maximum parsimony scores of the trees we used PAUP* 4.0 [15] For job management across the cluster and public laboratory machines, we used the Condor

software package [3]. We generated the rest of this software (a combination of C++ programs and Perl scripts) explicitly for these experiments.

4.2 Bounded Diameter Trees

We performed experiments on bounded diameter trees, and observe how the error rates increase as the number of taxa increases. We fix the diameters ($\max \lambda_{ij}$) at 0.01, 0.1 and 0.5, and the scaling factor at 4. For each diameter we generated 30 birth death trees (using r8s) of 10, 25, 50, 100, 200, 400 and 800 taxa. We then generated DNA sequences of sequence length 500 using seqgen, and observed how NJ and DCM-NJ+MP perform.

4.3 Varying Diameter Families of Trees

We also observed the performance of methods as the diameter ($\max \lambda_{ij}$) increases. For each method, NJ and DCM-NJ+MP, we look at their error rates for varying number of taxa, as the diameter increases.

We fixed the number of taxa, the scaling factor to 4 and generated 30 random birth death trees (using r8s) for diameters of 0.005, 0.01, 0.05, 0.1, 0.5, 1.0, 1.5 and 2. For each tree we then generated sequences of length 500 using seqgen and observed how NJ and DCM-NJ+MP perform.

We then repeated the above process for a different number of taxa. In total we looked at 10, 25, 50, 100, 200, 400 and 800 taxa. We have two separate graphs, one for NJ's performance and the other for DCM-NJ+MP's performance.

4.4 Results and Discussion

In Figure 1 we show how NJ and DCM-NJ+MP are affected by increasing the rate of evolution (i.e. the height). The x -axis is the maximum expected number of changes of a random site across the tree, and the y -axis is the false negative (FN) rate. We provide a curve for each number of taxa we explored, from 10 up to 800. The sequence length is fixed in this experiment to 500. Note that both NJ and DCM-NJ+MP have high errors for the lowest rates of evolution, and that at these low rates of evolution the error rates increase as n increases. This is because for these low rates of evolution, increasing the number of taxa makes the smallest edge length (i.e. f) decrease, and thus increases the sequence length needed to have enough changes on the short edges for them to be recoverable. As the rate of evolution increases, the error rates initially decrease for both methods, but eventually the error rates begin to increase again. This increase in error

occurs where the exponential portion of the convergence rate (i.e. where the sequence length depends exponentially on $\max \lambda_{ij}$) becomes significant. Note that where this happens is essentially the same for both methods – and that they perform equally well until that point. However, after this point, NJ’s performance is worse, compared to DCM-NJ+MP; furthermore, the error rate increases for NJ at each of the “large” diameters, as n increases, while DCM-NJ+MP’s error rate tends to not reflect the number of taxa nearly as much.

In Figure 2 we present a different way of looking at the data. In this figure, the x -axis is the number of taxa, and the y -axis is the FN rate, and there is a curve for each of the two methods. We show thus how increasing n (the number of taxa) while fixing the diameter of the tree affects the accuracy of the trees reconstructed. Note that at low rates of evolution (the left-most figure), the error rates for both methods increase with the number of taxa. At moderate rates of evolution (middle figure), error rates increase for both methods, but more so for NJ than for DCM-NJ+MP. Finally, at the higher rate of evolution (right figure), this trend continues, but the gap is even larger – in fact, DCM-NJ+MP’s error increase looks almost flat.

These experiments suggest strongly that except for low diameter situations, the DCM-NJ+MP method (and probably the other “fast converging” methods) will outperform the NJ method, especially for large numbers of taxa and high evolutionary rates.

5 Conclusion and Future Work

Our study provides convincing evidence that the NJ method is not likely to scale with increasing numbers of taxa, because the error rates grows significantly with the number of taxa and with the evolutionary rate. By contrast, the DCM-NJ+MP method shows a much more moderate increase in error for a fixed sequence length, as n increases (except for low rates of evolution, where the methods perform equally).

6 Acknowledgments

We want to thank the David and Lucile Packard Foundation (for a fellowship to Tandy Warnow), the National Science Foundation (for a POWRE grant to Katherine St. John), the Texas Institute for Computational and Applied Mathematics and the Center for Computational Biology at UT-Austin (for support of Katherine St. John), Doug Burger and Steve Keckler for the use of the SCOUT cluster at UT-Austin, and Patti Spencer and her staff for their help.

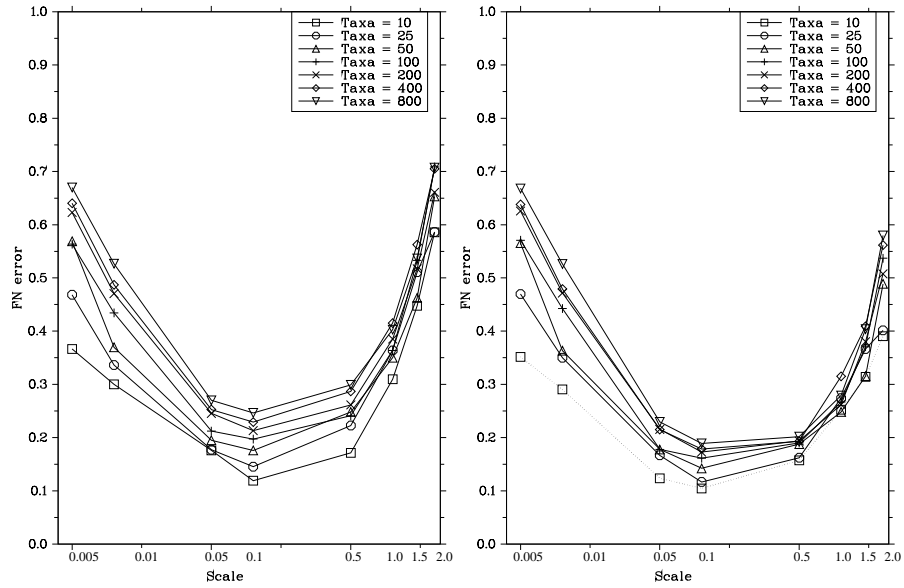


Fig. 1. NJ (left graph) and DCM-NJ+MP (right graph) error rates on birth death trees as the diameter (x-axis) grows. Sequence length fixed at 500 and scaling factor at 4

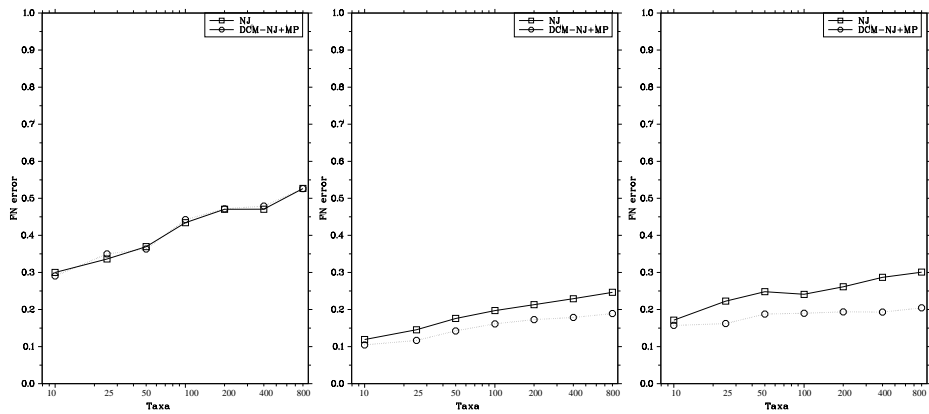


Fig. 2. NJ and DCM-NJ+MP Error rates on birth death trees as the number of taxa (x-axis) grows. Sequence length fixed at 500, diameter at 0.01 (left graph), 0.1 (middle graph), 0.5 (right graph) and scaling factor at 4.

References

1. K. Atteson. The performance of the neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25:251–278, 1999.
2. V. Berry and O. Gascuel. Inferring evolutionary trees with strong combinatorial evidence. In *Proc. 3rd Ann. Int'l Conf. Computing and Combinatorics COCOON97*, pages 111–123. Springer Verlag, 1997. in *LNCS 1276*.
3. Condor. Condor high throughput computing program, Copyright 1990-2001. Developed at the Computer Sciences Department of the University of Wisconsin; <http://www.cs.wisc.edu/condor/>.
4. M. Csűrös. Fast recovery of evolutionary trees with thousands of nodes. To appear in RECOMB 2001, 2001.
5. M. Csűrös and M. Y. Kao. Recovering evolutionary trees through harmonic greedy triplets. *Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA99)*, pages 261–270, 1999.
6. P.L. Erdos, Mike Steel, L. Székély, and Tandy Warnow. A few logs suffice to build almost all trees –I. *Random Structures and Algorithms*, 14:153–184, 1997.
7. P.L. Erdos, Mike Steel, L. Székély, and Tandy Warnow. A few logs suffice to build almost all trees –II. *Theor. Comp. Sci.*, 221:77–118, 1999.
8. J. P. Huelsenbeck and D. M. Hillis. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.*, 42:247–264, 1993.
9. D. Huson, S. Nettles, and Tandy Warnow. Disk-covering, a fast-converging method for phylogenetic tree reconstruction. *Comput. Biol.*, 6:369–386, 1999.
10. W. H. Li. *Molecular Evolution*. Sinauer, Massachusetts, 1997.
11. 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. ISMB01*. Oxford U. Press, 2001. to appear in *Bioinformatics*.
12. 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.
13. Michael Sanderson. available from <http://loco.ucdavis.edu/r8s/r8s.html>.
14. N. Sautou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.
15. D. L. Swofford. PAUP*: Phylogenetic analysis using parsimony (and other methods), 1996. Sinauer Associates, Underland, Massachusetts, Version 4.0.
16. Tandy Warnow, B. M. 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.