Performance Study of Phylogenetic Methods: (Unweighted) Quartet Methods and Neighbor-Joining

Katherine St. John* Tandy Warnow† Bernard M.E. Moret‡ Lisa Vawter§

Abstract

We present the results of a large-scale experimental study of quartet-based methods (quartet cleaning and puzzling) for phylogeny reconstruction. Our experiments include a broad range of problem sizes and evolutionary rates, and were carefully designed to yield statistically robust results despite the size of the sample space. We measure outcomes in terms of numbers of edges of the true tree correctly inferred by each method (true positives). Our results indicate that these quartet-based methods are much less accurate than the simple and efficient method of neighbor-joining, particularly for data composed of short- to medium-length sequences. We support our experimental findings by theoretical results that strongly suggest that quartet-cleaning methods are unlikely to yield accurate trees with less than exponentially long sequences. We conclude that a proposed reconstruction method should first be compared to the neighbor-joining method and further studied only if it offers a demonstrable practical advantage.

1 Introduction.

Reconstructing the evolutionary history of a group of taxa is a major research thrust in computational biology. The evolutionary process not only determines relationships among taxa, but also allows prediction of structural, physiological, and biochemical properties [5, 25]. Research on tree reconstruction has focused on reconstructing an evolutionary tree (phylogeny) under various optimization criteria. However, almost all optimization problems of interest to biologists are NP-hard (see [9] for a review), so that most biologists use heuristic methods or surrogate optimization criteria.

A popular family of phylogenetic heuristics is based on *quartets*. A quartet is an unrooted binary tree for a quadruple of taxa. For most optimization problems, it is possible to determine the optimal tree on a set of four leaves by analyzing all three possible trees. Quartet-based methods compute a quartet under an optimization criterion for each set of four taxa and combine the quartets into a tree on the full set of taxa. Because there are $\Theta(n^4)$ quartets, many quartet-based methods run in $\Omega(n^5)$ time, which is currently impractical for a hundred or more taxa.

How accurate are quartet-based methods at reconstructing phylogenetic trees? In biological applications, the true, historical tree is almost never known, which makes assessing the quality of phylogenetic reconstruction methods problematic (but see the study by Hillis *et al.* [8]). As a consequence, the method of choice for evaluating heuristics has been simulation [7]. In such a simulation, an ancestral biomolecular (DNA, RNA, or amino-acid) sequence is evolved along a "model" tree, producing a synthetic set of biomolecular sequences at the leaves. Phylogenetic reconstruction methods are then assessed based upon how accurately they reconstruct the model tree (the "true" tree). Biologists typically evaluate performance according to the topological accuracy of the reconstructed unrooted tree. The emphasis on topology is due to the biological interpretation of tree topology as the order of past speciation (or gene duplication) events, that is, relationships among species, genes, or other taxa. Topological accuracy is typically measured by the percentage of edges of the true tree found in the reconstructed tree (true positives).

Among distance-based methods (methods that transform input sequences into a distance matrix and then construct the tree from that distance matrix), none is more widely used by biologists than the *neighbor-joining (NJ)* method [22]. Not only is it quite fast ($O(n^3)$ for n taxa [24]), but experimental work has also shown that the trees NJ constructs are reasonably accurate, as long as the rate of evolution is neither too low nor too high. However, there is no comparative study of NJ and quartet-based methods.

*Dept. of Math & Computer Science, Lehman College and the Graduate Center, City U. of New York, stjohn@lehman.cuny.edu, supported in part by NSF grant 99-73874

†Dept. of Computer Science, U. of Texas at Austin, tandy@cs.utexas.edu, supported by NSF grant 94-57800 and by the David and Lucile Packard Foundation

‡Dept. of Computer Science, U. of New Mexico, moret@cs.unm.edu, supported in part by NSF grants CDA 95-03064 and ITR 00-81404

§Bioinformatics, SmithKline Beecham, Upper Merion, PA, lisa_vawter@sbphrd.com

We present the results of a detailed, large-scale experimental study of quartet-based methods and NJ under the Jukes-Cantor model of evolution [14]. Our results indicate that NJ always outperforms the quartet-based methods we examined, in terms of both accuracy and speed. We conclude that NJ, already the most popular distance-based method, should be used as a minimum standard in the assessment of phylogenetic methods: a proposed method should be compared with NJ and shown to provide a demonstrable advantage over it before that method is studied in depth. Finally, we present new theory about convergence rates of quartet-based methods which helps explain our observations.

2 Terminology and Review

2.1 The Jukes-Cantor model. The *Jukes-Cantor* model [14] is the simplest Markov model of biomolecular sequence evolution. In that model, a DNA sequence (a string over $\{A, C, T, G\}$) at the root evolves down a tree. The sites (i.e., the positions within the sequences) evolve independently and identically and the number of changes of each site on edge e is a Poisson random variable with expectation λ_e.

2.2 Measures of accuracy. Let T be the true tree, and let T' be an estimation of T, with both T and T' leaf-labelled by a set S of taxa. For each edge $e \in E(T)$, we define the bipartition π_e induced on S by the deletion of the edge e from T. The set $C(T) = \{\pi_e : e \in E(T)\}$ is called the "character encoding of T." Methods for reconstructing trees are evaluated according to the degree of topological accuracy obtained, by comparing the sets $C(T)$ and $C(T')$. The *true positives* are the edges $e \in E(T)$ obeying $\pi_e \in C(T) \cap C(T')$.

2.3 Statistical performance issues. Under the Jukes-Cantor model, a method M is *statistically consistent* if, for every model tree $(T, \{\lambda_e\})$ and every $\varepsilon > 0$, there is a sequence length k (which depends on M, T, λ, and ε) such that M recovers the true tree with probability at least $1 - \varepsilon$, when the method is given sequences generated on T of length at least k.

The sequence length required by a method is a significant aspect of its performance, because real data are of limited length (typically bounded by a few hundred to a few thousand nucleotides). Computational requirements are also important, but it may be possible to wait longer or use more powerful machines, whereas it is not possible to get longer sequences than exist in nature. Consequently, experimental and analytical studies have attempted to bound the sequence lengths required by different phylogenetic methods. The rate at which a method converges to 100% accuracy as a function of the sequence length is called the *convergence rate*.

2.4 Neighbor-joining. *Neighbor-Joining (NJ)* was formally described in 1987 [22] and has been a mainstay of phylogeny reconstruction among biologists ever since. NJ is an $O(n^3)$ algorithm that proceeds by repeatedly pairing two subtrees (at first, a pair of leaves; thereafter entire subtrees), replacing that pair in further computations with a single artificial taxon representing the subtree, thereby eventually returning a binary (fully resolved) tree. NJ is statistically consistent for the Jukes-Cantor model of evolution.

2.5 Quartet-based methods. A *quartet* is an unrooted binary tree on four taxa. A quartet thus induces a unique bipartition of the four taxa and can be denoted by that bipartition. If the taxa are $\{a, b, c, d\}$, we can use $\{ab|cd\}$ to denote the quartet that pairs a with b and c with d (see Figure 1). A quartet $\{ab|cd\}$

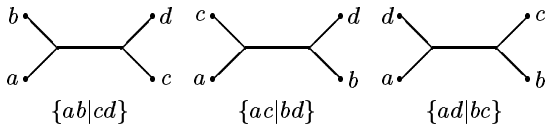


Figure 1: The three possible quartets on four taxa $\{a, b, c, d\}$ and their bipartition encodings.

agrees with a tree T if all four of its taxa are leaves of T and the path from a to b in T does not intersect the path from c to d in T. Equivalently, $\{ab|cd\}$ agrees with a tree if the subtree induced in T by the four taxa is the quartet itself. The quartet $\{ab|cd\}$ *is an error* with respect to the tree T if it does not agree with T. If $Q(T)$ denotes the set of all quartets that agree with T, then T is uniquely characterized by $Q(T)$ and can be reconstructed from T in polynomial time [6].

Quartet-based methods operate in two phases. In the first phase they construct a set Q of quartets on the different sets of four taxa. One popular approach is to use the computationally intensive, but statistically sophisticated method of "maximum likelihood estimation" (ML). In the second phase, they combine these quartets into a tree on the entire set of taxa. In practice, the input data are not of sufficient quality to ensure that all quartets are accurately inferred, so that quartet methods have to find ways of handling incorrect quartets. Most optimization problems related to tree reconstruction from quartets are NP-hard. An example of this is the *Maximum Quartet Compatibility* problem [13], which seeks a tree T for a given set Q of quartets such that $|Q(T') \cap Q|$ is maximized.

The methods studied in this paper have no performance guarantees with respect to the *Maximum Quartet*

Compatibility problem, although each of them is statistically consistent under the Jukes-Cantor model of evolution. However, with the exception of Quartet Puzzling, all quartet methods we examine do provide guarantees about the edges of the true tree that they reconstruct. These guarantees are expressed in terms of "quartet errors around an edge," a concept we now define.

Consider an edge e in the true tree T; its removal defines the bipartition $A|B$ on the leaves of S. Consider those sets of four leaves $\{a, a', b, b'\}$ with $\{a, a'\} \subseteq A$ and $\{b, b'\} \subseteq B$. A quartet $t \in Q$ is said to be an "error around e" if we have $t = \{ab|a'b'\}$ or $t = \{ab'|a'b\}$. Similarly, if T' is a proposed tree, and Q is a set of quartets, then $t \in T'$ is an error around edge $e \in E(T')$ if $t = \{ab|a'b'\}$ or $t = \{ab'|a'b\}$, while e defines the bipartition $A|B$.

Two of the methods we study, the Q^* method (also known as the Buneman method) and the Quartet Cleaning methods, can be described in terms of an explicit bound on the number of quartet errors around the edges they reconstruct. We begin with the Q^* method [3]. This method seeks the *maximally resolved* tree T' obeying $Q(T') \subseteq Q$. Therefore, there are *no quartet errors* around any edge in the tree T'. This tree always exists, since the star tree trivially satisfies the constraint on any set of quartets. The Q^* tree is unique and can be constructed in polynomial time; by design, however, the Q^* method is conservative and generally produces very unresolved trees [11].

Quartet-Cleaning (QC) methods [2, 4, 13] have explicit bounds on the number of quartet errors around each reconstructed edge e. These error bounds have the form $m\sqrt{q_e}$, where q_e is the number of quartets around edge e and m is a small constant. Thus, the Q^* method is a cleaning method in which $m = 0$. The *global cleaning* method sets $m = 1$ and the *local cleaning* method sets $m = \frac{1}{2}$; these methods are guaranteed to recover every edge of the true tree for which Q contains a small enough number of quartet errors. The *hypercleaning* method allows m to be an arbitrary integer and thus has the potential to recover more edges. However, its running time is very high—proportional to $n^7 \cdot m^{4m+2}$—so that it is impractical for m larger than 5. We investigated whether these error bounds are sufficiently large to avoid the resolution problems encountered by the Q^* method and, in particular, whether the performance of QC methods scales with increasing problem size.

The final quartet-based method we examined is the best known and the most frequently used by biologists [17, 21, 15]: the *Quartet-Puzzling (QP)* method [23]. This heuristic computes quartets using ML and then uses a greedy strategy to construct a tree on which many input quartets are in agreement. QP uses an arbitrary ordering of taxa, constructs the optimal quartet on the first four, then inserts each successive taxon in turn, attaching the new leaf to an edge of the current tree so as to optimize a quartet-based score. Because the input ordering of taxa is pertinent, QP uses a large number of random input orderings and computes the *majority consensus* of all trees found. The majority consensus is the tree that contains all bipartitions that appear in more than half of the trees in the set and is a well-known consensus method among biologists. Thus, QP implicitly seeks to return a tree in which every edge is "well-supported," in the sense that each edge appears in more than half the trees obtained during the algorithm and thus has (presumably) many supporting quartets.

2.6 Previous experimental studies of quartet methods. Berry *et al.* conducted experimental studies of various QC methods [2, 4]. They evolved sequences on model trees, compared the quartets inferred by various methods with the quartets of the true tree, and determined which edges of the model tree could be reconstructed by their QC method. They varied evolutionary rates and sequence lengths, but only examined trees with ten taxa. Their results showed that QC methods, especially hypercleaning, outperform the $Q*$ method with respect to true positives. (Of course, by design, the QC methods cannot fail to recover an edge that is recovered by the Q^* method. So what is noteworthy in the experiments is that the QC methods *did* succeed in obtaining additional edges.) Because the dataset sizes used in these experiments are quite small (only 10 taxa), these results may not generalize to larger numbers of taxa; indeed, the theoretical bounds we derive on the convergence rate of QC methods suggest that performance on larger n may be poor. Finally, no comparison was made between QC methods and NJ or other tree reconstruction methods.

3 Theoretical Bounds on the Convergence Rates.

We begin with the known upper bounds on the convergence rates of NJ and the Q^* method. Surprisingly, these are identical [1, 6], although experimental studies strongly suggest that NJ obtains accurate reconstructions of trees from shorter sequences than Q^* throughout the parameter space of Jukes-Cantor trees [11].

THEOREM 3.1. *Let $f, g, \varepsilon > 0$ be arbitrary constants with $f < g$. Denote by $B(S)$ the tree reconstructed on S by the Q^* method and by $NJ(S)$ the tree reconstructed by NJ. There is a constant $c > 0$ such that, for all Jukes-Cantor trees $(T, \{\lambda_e\})$ on n leaves with $0 < f \leq \lambda_e \leq g < \infty$ for all $e \in E(T)$, and for a set S of sequences generated randomly on T,*

$$Pr[B(S) = NJ(S) = T] > 1 - \varepsilon$$

if the sequence length exceeds $c \log n \cdot e^{O(g \cdot diam(T))}$, *where* $diam(T)$ *is the length of a longest path in* T.

Because the diameter of an n-leaf tree can be as much as $n-1$ (and is typically $\Omega(\sqrt{n})$ [6]), Theorem 3.1 shows that the Q^* and NJ methods will converge from sequences that grow exponentially in n. While Theorem 3.1 provides only an upper bound, earlier experimental work shows that the Q^* method performs quite poorly when g and $diam(T)$ are both large [12], and that NJ is also affected, although less severely [11].

We now consider the convergence rate of the QC methods. Because the error bound used in QC methods is a multiple of $\sqrt{q_e}$, the ratio of permitted errors to the number of quartets around an edge is $m/\sqrt{q_e}$. Because we have $q_e = \Omega(n^2)$ and because m is a small constant, this ratio rapidly approaches 0 as the number of taxa increases. For example, consider an edge in a 50 taxon tree producing a 20 : 30 split. The number of quartets around this edge is $82,650$, so that the bound for local cleaning is only 144; hypercleaning with $m = 5$ brings this bound up to 1440. Thus, for 50 taxa, even hypercleaning has an error tolerance on some edges that is less than 2% of the total number of quartets for this edge.

The sensitivity of QC methods to errors suggests that, for large n, QC methods will be close in performance to the Q^* method. As noted earlier, the convergence rate of the Q^* method is bounded from above by a function that grows exponentially in n, so that the Q^* method is impractical. If cleaning methods tend to perform only as well as the Q^* method for large n, then they will not scale well.

Consider therefore a hypothetical cleaning method we will call *HypoClean*. This method is guaranteed to recover an edge e if the number of quartet errors around e is at most one third of the quartets around the edge—a much more generous bound than that used in local cleaning. In the following theorem we establish a bound on the sequence length that suffices for *HypoClean* to be accurate on a random Jukes-Cantor tree.

We require the following lemma.

LEMMA 3.1. *The median diameter of all* $(2n-5)!!$ *unrooted, leaf-labelled, binary trees on* n *leaves is* $\Theta(\sqrt{n})$.

Proof. Penny and Steel [19] gave formulas for the distribution of interleaf distances in such trees under the assumption that all $(2n-5)!!$ such trees are equally likely, obtaining

$$\mu(D) = 2^{2n} / \binom{2n}{n} = \Theta(\sqrt{\pi \cdot n})$$

and

$$\sigma^2(D) = 4n - 6 - \mu(D) - \mu^2(D) = \Theta((4-\pi) \cdot n)$$

Because any nondegenerate distribution must have its median within $[\mu - \sigma, \mu + \sigma]$, our conclusion follows.

THEOREM 3.2. *Let* $f, g, \varepsilon > 0$ *be arbitrary constants with* $f < g$ *and denote by* $HC(S)$ *the tree reconstructed on* S *by the* HypoClean *method. Then there is a constant* c *such that, for random Jukes-Cantor trees,* $(T, \{\lambda_e\})$ *with* $0 < f \leq \lambda_e \leq g < \infty$ *for all* $e \in E(T)$ *and for a set* S *of* n *sequences generated randomly on* T, *we have* $Pr[HC(S) = T] > 1 - \varepsilon$ *whenever the sequence length exceeds* $c \log n \cdot e^{O(g \cdot \sqrt{n})}$.

Proof. Theorem 3.1 shows that quartets of low diameter are more easily reconstructed from short sequences than are those quartets of high diameter. Assume that we can correctly reconstruct the "smallest-diameter" half of the quartets with high probability—we simply guess the remaining quartets. We will then correctly reconstruct 2/3 of the quartets with high probability. What sequence length is required for this? Solving the smaller half of the quartets is no easier than solving the median-diameter quartets. By Theorem 3.1, this latter task is achieved with high probability when the sequence length is at least $O(c \log n \cdot e^{O(g \cdot md(T))})$, where $md(T)$ is the median diameter of T. By Lemma 3.1, this quantity is $\Theta(\sqrt{n})$. Therefore, the sequence length that suffices to reconstruct the true tree with high probability using the *HypoClean* method is $O(c \log n \cdot e^{O(g \cdot \sqrt{n})})$.

Thus all cleaning methods have the same upper bound on their rate of convergence, indicating that these methods may not scale well.

4 Experimental Design.

4.1 Overview. We used Jukes-Cantor model trees with varying numbers of taxa and rates of evolution to generate a large number of synthetic datasets of varying lengths. For each dataset generated, we computed the NJ and QP trees on the entire dataset and two sets of quartets, one based upon ML, Q_{ML}, and one based upon NJ, Q_{NJ}. We then applied various cleaning methods to each of the sets Q_{ML} and Q_{NJ}. We compared quartets of Q_{ML}, of Q_{NJ}, and of the reconstructed trees, as well as the reconstructed trees themselves, against the model tree for accuracy.

4.2 Model trees. We randomly generated model tree topologies from the uniform distribution on binary leaf-labelled trees. For each edge of each tree topology, we generated a random number (from the uniform distribution) between 1 and 1000 and used that number as the "length" of the edge. We then scaled each such "base" model tree by a multiplicative factor, ranging from 10^{-7} to 10^{-3}. This process produces Jukes-Cantor trees with λ_e values ranging from a minimum of 10^{-7} to a maximum of 1. We generated random DNA sequences for the root and used the program `Seq-Gen`

[20] to evolve these sequences down the tree under the Jukes-Cantor model of evolution, thus producing sets of sequences at the leaves, our synthetic datasets.

4.3 Statistical considerations. Because the number of distinct unrooted, leaf-labelled trees on n leaves is $(2n-5)!!$ and because our input space is further expanded by the choice of evolutionary rates, it is not possible to take a fair sample of the entire input space. In order to obtain statistically robust results, we followed the advice of McGeoch [16] and Moret [18] and used a number of *runs*, each composed of a number of *trials* (a trial is a single comparison), computed the mean outcome for each run, and studied the mean and standard deviation over the runs of these events. This approach is preferable to using the same total number of samples in a single run, because each of the runs is an independent pseudorandom stream. With this method, one can obtain estimates of the mean that are closely clustered around the true value, even if the pseudorandom generator is not perfect.

4.4 Parameter space. A critical parameter of our study, one that has not been explored in most prior studies, is the number of input taxa. Previous experimental studies have often been limited to a small number of taxa due to computational problems. However, to resolve phylogenetic trees of interest to biologists, algorithms must scale reasonably, both in terms of topological accuracy and running time, to problems of the size that biologists typically study (20–200 taxa), as well as those they would like to address (a few hundred to several thousand taxa).

Because of dedicated use of two multiprocessor clusters, we were able to run our test suite for 5, 10, 20, and 40 taxa (full quartet-based methods remain impractical, at least in terms of experimental studies, for large numbers of taxa). Our tests included a selection of eight expected evolutionary rates, from 5×10^{-5} to 5×10^{-1} per tree edge. For each evolutionary rate and problem size, we generated a total of 100 topologies, grouped into 10 runs of 10 trials. All tests were conducted for four sequence lengths: 500, 2,000, 8,000, and 32,000 (we note that sequence lengths above 1,000 are considered long and those above 5,000 extremely long; thus our study explores longer sequence lengths than are usually encountered in practice). In all, our study used 16,000 datasets and required many months of computation on the two clusters.

4.5 Algorithms. We tested four different phylogenetic reconstruction methods: NJ, local quartet-cleaning for quartets based on NJ, local quartet-cleaning for quartets based on ML, and QP. The code for QP is `TREE-PUZZLE`, available from their authors at `www.tree-puzzle.de`. We modified the QP source only by removing its interactive interface; all other code is our own. For quartet-cleaning, our accuracy measurements were made directly on an edge-by-edge basis—the actual tree was not reconstructed; in contrast, QP and NJ actually reconstruct a tree. We ran all four algorithms sequentially on a single set of sequences for one trial, stored all data that was generated, then proceeded to the next trial, so that each of the algorithms was run on exactly the same data.

4.6 Measurements. Our focus in this study is the accuracy of solutions generated by the various tree reconstruction methods. Because most methods are time-consuming, the running time is briefly addressed; our aim was not to fine-tune implementations, but simply to obtain a rough idea of which methods can be run in a reasonable amount of time on a conventional machine for realistic datasets. We compare running times as gathered on our platforms, all of which are 450MHz Pentium III machines running Linux.

To assess topological accuracy, we measured the number of true positives (edges of the true tree that appear in the reconstructed tree). For cleaning methods, we measured these values before and after cleaning. For each run of 10 trials, we retained only the mean values. Our results are composed of the means for each set of 10 runs.

5 Experimental Results.

Space limitations preclude us from showing data on variance. Suffice it to say that, except for runs on 5 taxa, the standard deviation we observed remained consistent at 1–2% of the mean; with 5 taxa, standard deviations were larger, reaching 10–15% of the mean. In all of our figures, QCNJ and QCML denote quartet-cleaning of quartets derived by local NJ and by ML, respectively.

5.1 Estimating quartets. The technique used to construct the set Q of quartets provided to quartet-based methods can have a significant impact on the performance of these methods. The phylogenetics community has generally expected that ML would produce more accurate quartets than NJ. We therefore compared ML and NJ in terms of the quartet sets, Q_{NJ} and Q_{ML}, that they computed. As a reference point, we also examined how global NJ performed in terms of the trees it induced on each fourtuple of leaves from the NJ tree. Figure 2 shows the proportion of true positives in each of the sets of quartets.

The relative performance of local NJ and local ML

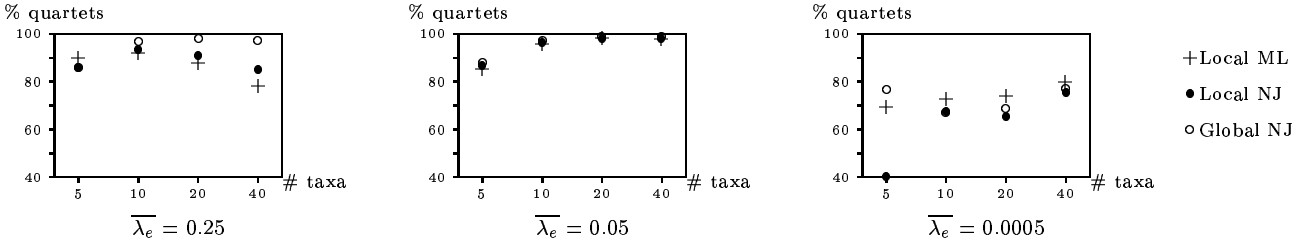


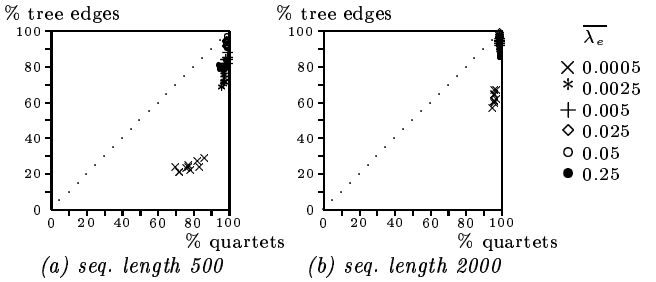
Figure 2: Percentages of quartets computed by local ML, local NJ and induced by global NJ that agree with the model tree for various numbers of taxa and a sequence length of 500.

(NJ and ML applied to each quadruple of leaves to estimate the quartets) is interesting. At the highest rates of evolution (except for five-taxon trees) local NJ slightly outperforms local ML, but this gap increases with increasing numbers of taxa. At the second highest rate of evolution they are indistinguishable up to 40 taxa. However, at the lowest rate of evolution, local ML slightly outperforms local NJ, although the gap decreases with increasing numbers of taxa. In general, this suggests that quartets with large evolutionary distances are more accurately inferred by NJ than by ML. More generally, neither ML nor NJ dominates the other in terms of accuracy; each has a range in which it yields slightly better quartet estimations.

A comparison between these sets of quartets to the quartets obtained by using global NJ (i.e., the quartets induced by the NJ tree) is also interesting. At the lowest rate of evolution (except for five-taxon trees), local ML is superior to global NJ, and both are superior to local NJ; however, the gap between the three ways of computing quartet trees narrows with increasing number of taxa. At the middle rate, the methods are indistinguishable (up to 40 taxa), while at the highest rate, global NJ is clearly superior, and the gap between global NJ as a quartet method and the two other quartet methods increases with increasing numbers of taxa. Thus, for high rates of evolution (and potentially for all large enough trees), the best quartet estimator may simply be global NJ—i.e., compute the NJ tree and use its quartets! (We note that this way of defining quartet trees is not suited to quartet methods, as they would then combine the quartet trees back into the NJ tree, and nothing would be gained.)

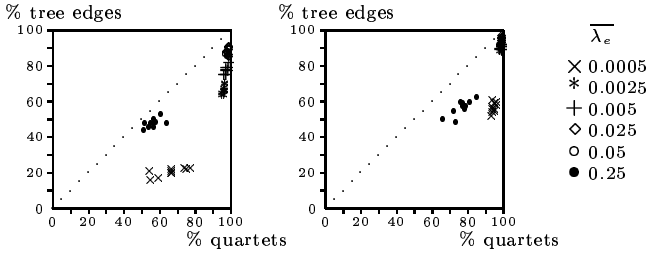
In terms of the quality of the quartets obtained, the best accuracy was obtained at the second highest rate of evolution. At the lowest rate of evolution, only 1 in 2000 sites changes on average, so that, for a sequence length of 500, roughly 25% of the edges have no changes on them. Thus, although it may be possible to guess an edge accurately, the best possible reconstruction at the lowest rate will only yield about 75% of the edges—approximately what the best performing method (local NJ) obtains. At the higher rates of evolution we examined, all methods increased in accuracy. At the highest rate the accuracy starts to decrease with more than 10 taxa. The decrease in accuracy with increasing numbers of taxa at the highest rate of evolution is predicted by theory (if only for information-theoretic reasons); hence, even for the lower rates of evolution, as the number of taxa increases, the accuracy of the quartet estimations should decrease.

5.2 Two measures of accuracy: quartets and edges. Although the standard measure of accuracy is the number of true edges in the reconstructed tree, the percentage of correctly inferred quartets has also been used as a surrogate [4]. However, correlation between correct quartets and edges of the true tree returned by a method has not been shown. We address this issue by examining the performance of QP and global NJ with respect to both criteria. Figures 3 and 4 make it clear that topological accuracy is a more demanding criterion than quartet accuracy, and should therefore be used to assess performance of phylogenetic reconstruction methods. Both NJ and QP can return trees with only 20% of the edges correct from a set of quartets that is 80% correct. Worse yet, both methods, except when the percentage of correct quartets is close to 100%, can return fewer than 80% of the true tree edges (in the case of QP, some such trees had only 60% of the true



(a) seq. length 500 *(b) seq. length 2000*

Figure 3: Percent of true tree edges recovered by global NJ for various $\overline{\lambda_e}$ as a function of the percentage of correct induced quartets for 40 taxa and two sequence lengths



(a) seq. length 500 *(b) seq. length 2000*

Figure 4: Percent of true tree edges recovered by QP for various $\overline{\lambda_e}$ as a function of the percentage of correct induced quartets for 40 taxa and two sequence lengths

tree edges). Because failure to obtain at least 90 or 95% of the edges can be unacceptable to systematists, quartet-based measures of accuracy are not acceptable surrogates for true tree edges.

5.3 Sensitivity to input quality. Methods that operate by estimating quartets and then combining them into a single tree can be greatly affected by the quality of the input quartets. Figure 5 shows how QC methods respond to input quality. QC methods, as well as the other quartet methods we study, require a larger fraction of correct input quartets than the fraction of true tree edges that they return.

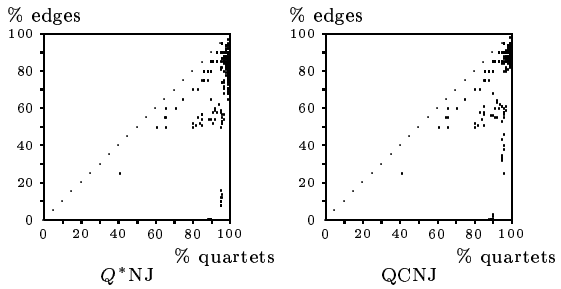


Figure 5: Percentage of correct input NJ quartets *vs.* true tree edges for Q^*NJ and QCNJ for sequence length fixed at 2000, with each graph showing runs for all numbers of taxa and all average edge lengths.

5.4 Scaling of methods with increasing numbers of taxa. Theory predicts that the accuracy of methods will degrade as the number of taxa increases while sequence length and average edge length (the expected number of changes for a random site on each edge) are held fixed. Space limitations force us to show only a sample of our results. Figure 6 shows the topological accuracy achieved by all six methods as a function of the number of taxa for a sequence length of 500 and for three different average edge lengths. Figure 7 shows the same set of results for a sequence length of 2000. All methods decrease in accuracy as the number of taxa increases, even though both NJ and QP show an initial increase. QC provides a distinct improvement over the Q^* method, whether the quartets are computed using ML or local NJ. QCML and QCNJ are very close in performance, although QCNJ slightly outperforms QCML; similarly Q^*NJ slightly outperforms Q^*ML. Of the five quartet methods, QP is the best throughout the range of parameters studied, but NJ completely dominates it.

5.5 A comparison between Q^* and QC. QC can be seen as an improvement to the Q^* method, because Q^* does not permit errors around any reconstructed edges, while QC reconstructs every edge around which there is a bounded number of errors. In Figures 6 and 7 we showed performance for different rates of evolution as the number of taxa varies, which gives evidence that QC methods return additional true edges under many conditions. In Figure 8 we explore the relative improvement in edge recovery obtained on local NJ or ML quartets by using a QC method instead of the Q^* method. Curiously, the improvement obtained in terms of quartet accuracy is less satisfactory, never averaging more than one percent. QC provides the most improvement when almost all input quartets are correct; indeed, this is what the theory about QC suggests. In particular, the most improvement occurs at a high rate of evolution—not our highest rate, but our second highest rate, when the error rate in input quartets is also lowest.

5.6 Rates of evolution and topological accuracy. Although sequence length and rate of evolution have a strong effect on the absolute performance of phylogenetic methods, the relative performance of NJ, QP, and QCNJ is constant throughout our experiments: NJ is the best followed by QP, and then by QCNJ. Figure 9 presents data for 40 taxa at three different rates of evolution, for sequence lengths varying from 500 (a typical length) up to 32,000 (a quite large length). Note that all methods increase in accuracy with increasing sequence length (as expected since all methods are statistically consistent under the Jukes-Cantor model).

6 Discussion.

6.1 Quality of quartets. The technique used to construct the set Q of input quartets provided to quartet-based methods can have a significant impact on the performance of these methods. The phylogenetics community has generally expected that ML would produce more accurate quartets than other methods. However, in our studies, neither ML nor NJ dominates the other as a quartet estimator; instead, ML outperforms NJ only for the lowest rates of evolution, whereas NJ clearly outperforms ML for higher rates. Because

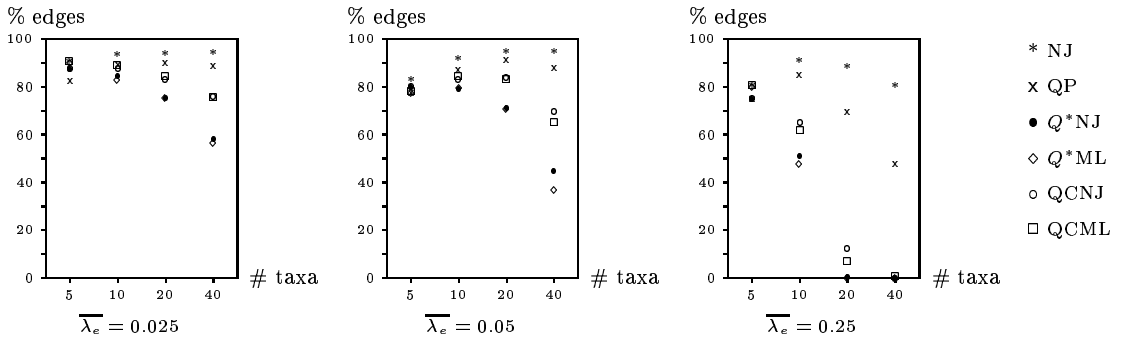


Figure 6: Number of taxa *vs.* percentage of edges correct for sequences of length 500 and various $\overline{\lambda_e}$

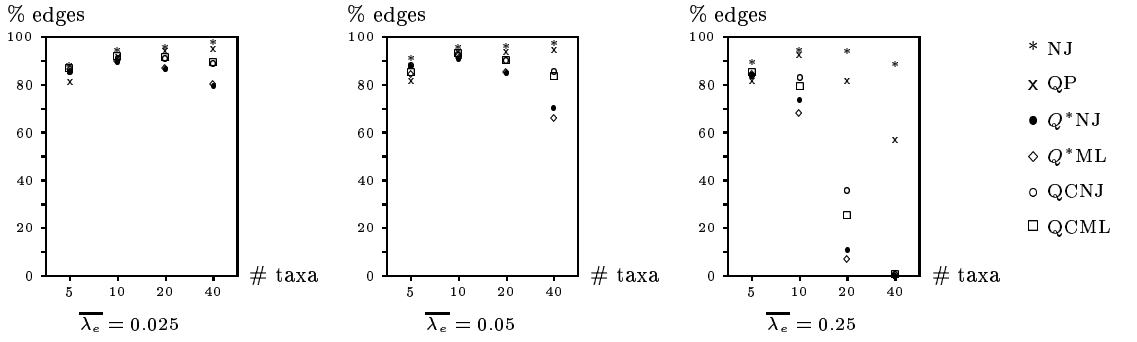


Figure 7: Number of taxa *vs.* percentage of edges correct for sequences of length 2000 and various $\overline{\lambda_e}$

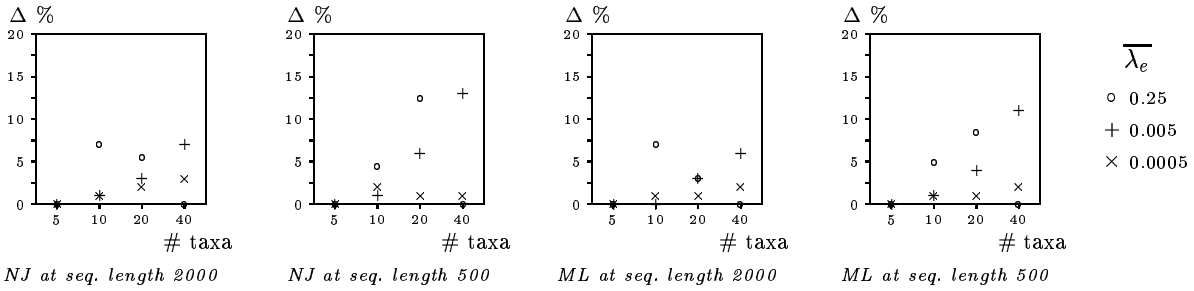


Figure 8: QC *vs.* Q^*: cleaning-induced improvement for NJ and ML in the percentage of tree edges that agree with the model tree.

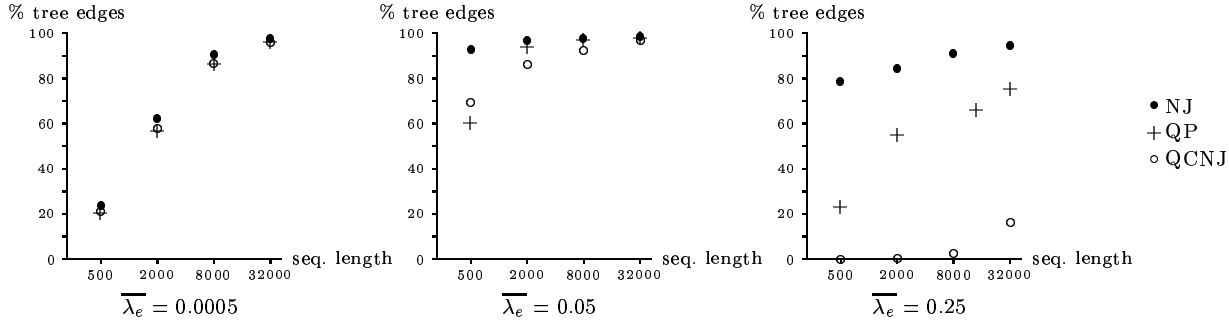


Figure 9: Accuracy of various methods as a function of sequence length for 40 taxa.

our observations differ from the received wisdom in the field, we offer the following possible explanation. In earlier studies [10], the performance of ML and NJ as quartet estimators was studied by explicitly simulating evolution on four-taxon trees. Here, we have simulated evolution on larger trees and then looked at the quartets defined by these larger trees. Good performance on quartets drawn from a large tree is not the same as good performance on quartets drawn from a much smaller sample space. While it is possible to sample four-taxon model trees so as to produce the same kind of quartets we gave as input to our methods, the studies in [10] did not use such a sampling strategy.

6.2 Robustness of quartet methods to quartet errors. How robust are quartet-based methods with respect to errors in Q? The Q^* method is the least robust. QC methods provide some error tolerance, sufficient to recover additional true edges even under high rates of evolution and for moderate numbers of taxa. However, both of these methods are inferior to QP in terms of error tolerance, even though QP also fails to get a good estimation of the true tree when Q has less than about 95% accuracy (for $n = 40$). Finally, in our experiments, NJ was always at least as accurate as QP and nearly always much better. Thus, the reason quartet methods fail to recover good trees is *not* that the input distance matrix is too noisy for any method to recover a good estimate of the true tree.

6.3 Running times. NJ was clearly the fastest method tested. QC and QP methods must compute all $\Theta(n^4)$ quartets and hence must take $\Omega(n^4)$ time. ML-based methods also construct quartets through expensive estimation methods, the running time of which increases sharply with increasing sequence length. Thus QCML and QP were by far the slowest of the methods tested, slow enough that running them on more than a hundred taxa appears infeasible at present. With default settings, QP takes more than 200 days of computation to analyze ten runs of ten trials each for a single set of parameters on 80 taxa with a sequence length of 500. In contrast, NJ dispatches the same analysis in about 30 minutes. For this reason, we have omitted any comparative results for the 80-taxon datasets. Our future research will examine these larger datasets.

6.4 Comparison between methods. Our experiments clearly establish a linear order of accuracy for the methods we studied (except under very low rates of evolution): NJ (applied globally) is the preferred method, with QP as a close second, the QC methods significantly behind QP, and the Q^* methods somewhat behind the QC methods. The particular technique used to infer quartets also has an influence on the quality of the trees obtained by the quartet methods, with QCNJ often better than QCML and Q^*NJ often better than Q^*ML (at least for large enough trees with moderate to high rates of evolution).

7 Conclusions and Open Questions.

Why does NJ outperform the quartet methods throughout the parameter space we examined (except on some five-taxon trees)? A possible explanation is that our upper bounds on the convergence rates of cleaning methods (and hence also of the Q^* method) are reasonably tight, whereas the performance of NJ clearly does not match its current best upper bound. Indeed, the sharp degradation in accuracy that we see in cleaning methods with increasing numbers of taxa suggests that our bounds are tight. In contrast, QP and NJ degrade far more gracefully, and only when the rate of evolution is close to saturation. Thus, the true convergence rates of QP and NJ may not be quite as poor as that of the cleaning methods.

If NJ and QP have true convergence rates that are lower than the current upper bound, what is the reason? The good performance of QP as a quartet method does not seem to result from its use of ML-based quartets, since by that reasoning QCML should demonstrate a comparable improvement over QCNJ (which it does not). Thus the reason for the better behavior of QP must lie in the manner in which it combines quartets. We suspect that the issue is partly that the Q^* and QC methods place too stringent a requirement on the edges; by comparison the QP method places no absolute restriction. Thus, we suspect that the ability of NJ and QP to handle distance data that have significant errors lies in the specific techniques each uses to construct trees and the fact that neither places excessively strict bounds on errors.

This in itself may help explain why QP outperforms the other quartet methods we studied, but it does not explain why NJ outperforms QP. Our conjecture is that methods that operate by combining quartets do not make use of all available information: we suggest that quartet-based methods may be impeded by their very structure, in having to decide the tree based on quadruples of taxa, without reference to the other taxa.

These observations suggest that quartet methods, if they are to be competitive with global NJ, cannot afford to place absolute error bounds, but need to be flexible in combining quartets into a single tree on the full set of taxa. Because of the lack of correlation between quartet accuracy and edge accuracy, seeking to realize the maximum number of quartets may not produce the

best trees either. Therefore, to design a quartet method with good performance (reaching or improving upon NJ's performance), seems to require both flexibility and greater sophistication than the current quartet methods utilize. Finally, quartet methods based on maximum likelihood might outperform global NJ when the data are generated in a model (other than the Jukes-Cantor model) for which statistically consistent distance-estimation techniques do not exist, yet for which maximum likelihood remains statistically consistent. In these conditions, QP in particular may outperform NJ.

We conclude with the following comments about algorithm design and performance studies in phylogenetics. From the perspective of experimental performance studies and algorithm design, NJ should be regarded as a universal lowest common denominator in phylogeny reconstruction algorithms. Its speed makes it easy to use under all circumstances; its topological accuracy makes it an acceptable starting point for tree reconstruction in biological practice. We suggest that a proposed method should be compared with NJ and abandoned if it does not offer a demonstrable advantage over NJ for substantial subproblem families.

Acknowledgements: The authors wish to thank the Department of Computer Science at UNM for hosting KS, TW, and LV in Summer 2000.

References

[1] Atteson, K., "The performance of the neighbor-joining method of phylogeny reconstruction," *Algorithmica* **25**, 2/3 (1999), 251–278.

[2] Berry, V., Bryant, D., Jiang, T., Kearney, P., Li, M., Wareham, T., and Zhang, H., "A practical algorithm for recovering the best supported edges of an evolutionary tree," preprint (2000).

[3] Berry, V., and Gascuel, O., "Inferring evolutionary trees with strong combinatorial evidence," to appear in *Theor. Comp. Sci.*

[4] Berry, V., Jiang, T., Kearney, P., Li, M., and Wareham, T., "Quartet cleaning: improved algorithms and simulations," *Proc. Europ. Symp. Algs. (ESA99)*, LNCS **1643**, 313–324.

[5] Chambers, J.K., *et al.*, "A G protein-coupled receptor for Uridine 5'-diphosphoglucose (UDP-glucose)," *J. Biol. Chem.* **275** (2000), 10767–10771.

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

[7] Hillis, D.M., "Approaches for assessing phylogenetic accuracy," *Syst. Biol.* **44** (1995), 3–19.

[8] Hillis, D.M., Bull, J.J., White, M.E., Badgett, M.R., and Molineux, I.J., "Experimental phylogenies: generation of a known phylogeny," *Science* **255** (1992), 589–592.

[9] Hillis, D.M., Moritz, C., and Mable, B. *Molecular Systematics.* Sinauer Pub., Boston, 1996.

[10] Huelsenbeck, J., and Hillis, D. "Success of phylogenetic methods in the four-taxon case." *Syst. Bio.* **42 (3)**: (1993), 247–264.

[11] Huson, D., Nettles, S., Rice, K., Warnow, T., and Yooseph, S., "The hybrid tree reconstruction method," *ACM J. Experimental Algorithms* **4**, 5 (1999), `www.jea.acm.org/1999/HusonHybrid/`.

[12] Huson, D., Nettles, S., and Warnow, T., "Disk-covering, a fast converging method for phylogenetic tree reconstruction," *J. Comp. Biol.* **6** (1999), 369–386.

[13] Jiang, T., Kearney, P.E., and Li. M., "A polynomial-time approximation scheme for inferring evolutionary trees from quartet topologies and its application," preprint (2000).

[14] Jukes, T.H., and Cantor, C. *Mammalian Protein Metabolism.* Academic Press, NY (1969), 21–132.

[15] Keeling, P.J., Luker, M.A., and Palmer, J.D., "Evidence from beta-tubulin phylogeny that microsporidia evolved from within the Fungi," *Mol. Biol. & Evol.* **17** (2000), 23–31.

[16] McGeoch, C.C., "Analyzing algorithms by simulation: variance reduction techniques and simulation speedups," *ACM Comp. Surveys* **24** (1992), 195–212.

[17] Mishof, B., Anderson, C.L., and Hadrys, H., "A phylogeny of the damselfly genus *Calopteryx* (Odonata) using mitochondrial 16s rDNA markers," *Molec. Phylog. & Evol.* **15** (2000), 5–14.

[18] Moret, B.M.E., "Towards a discipline of experimental algorithmics," to appear in *Proc. 5th DIMACS Challenge*, available at `www.cs.unm.edu/m̃oret/dimacs.ps`.

[19] Penny, D., and Steel, M.A., "Distributions of tree comparison metrics—some new results," *Syst. Biol.* **42** (1993), 126–141.

[20] Rambaut, A., and Grassly, N.C., "Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees," *Comp. Applic. Biosci.* **13** (1997), 235–238.

[21] Rodrigues-Trelles, F., Alarcon, L., and Fontdevila, A., "Molecular evolution and phylogeny of the *buzzatii* complex (*D. repleta* group): a maximum likelihood approach," *Mol. Biol. & Evol.* **17** (2000), 1112–1122.

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

[23] Strimmer, K., and von Haeseler, A., "Quartet puzzling: a maximum likelihood method for reconstructing tree topologies," *Mol. Biol. & Evol.* **13** (1996), 964–969.

[24] Studier, J.A., and Keppler, K.J., "A note on the neighbor-joining method of Saitou and Nei," *Mol. Biol. & Evol.* **5** (1981), 729–731.

[25] Szekeres, P.G., *et al.*, "Neuromedin U is a potent agonist at the orphan G protein coupled receptor FM3," *J. Biol. Chem.* **275** (2000), 20247–20250.