The Short Quartet Method

Péter L. Erdős
Kenneth Rice
Michael A. Steel
László A. Székely
and Tandy J. Warnow

[1] Mathematical Institute of the Hungarian Academy of Sciences. E-mail: elp@math-inst.hu
[2] Department of Biology, University of Pennsylvania. Email: krice@saul.cis.upenn.edu
[3] Biomathematics Research Centre, University of Canterbury. E-mail: m.steel@math.canterbury.ac.nz
[4] Department of Mathematics, University of South Carolina. E-mail: laszlo@math.sc.edu
[5] Department of Computer and Information Science, University of Pennsylvania. E-mail: tandy@central.cis.upenn.edu.

Abstract. Reconstructing phylogenetic (evolutionary) trees is a major research problem in biology, but unfortunately the current methods are either inconsistent somewhere in the parameter space (and hence do not reconstruct the tree even given unboundedly long sequences), have poor statistical power (and hence require extremely long sequences on large or highly divergent trees), or have computational requirements that are excessive. We describe in this paper a new method, which we call the *Short Quartet Method*, for inferring evolutionary trees. The Short Quartet Method has great statistical power, is provably consistent throughout the parameter space, and uses only polynomial time. We present the results of experimental studies based upon simulations of sequence evolution that demonstrate its greater statistical power than neighbor-joining [33], perhaps the most popular method for phylogenetic tree inference among molecular biologists.

1 Introduction

The study of evolution is a fundamental problem in Biology, and advances in this area are of tremendous value to biomedical sciences. For example, understanding the evolutionary history of a set of DNA sequences can answer such questions as whether humans first originated in Africa (the controversial *African-Eve hypothesis*) [42], assist in the design of drugs to cure or control diseases, potentially determine the origins of life, as well as helping to answer immediate biomedical questions such as whether a particular Florida dentist infected his patients with HIV. Yet many of these questions remain essentially unanswered because the problem of inferring accurate evolutionary trees is extraordinarily difficult. Two of the critical issues that make evolutionary tree reconstruction difficult are that

every method, no matter how well suited, has on each tree an implicit sequence length it needs in order to be accurate with high probability, as well as computational requirements that are needed in order to use them method (many methods are, for example, attempts to solve NP-hard optimization problems, and hence can take a very long time on some instances). These two issues in particular make the reconstruction of very large trees containing widely divergent pairs of sequences enormously difficult. Consequently, many in systematic biology have considered the reconstruction of large evolutionary trees to be beyond the current capabilities of today's software and hardware.

In this paper, we describe a new method for reconstructing evolutionary trees which we call the *Short Quartet Method*. Our method is surprisingly simple, and has sufficient statistical power to reconstruct even very large and divergent trees from sequences that are realistically bounded in length. Furthermore, the method uses only polynomial time. We briefly describe the Short Quartet method and present the results of an experimental study in which we compare it to neighbor-joining, the most popular method for tree reconstruction among molecular biologists.

2 Phylogeny Estimation: the Current State

2.1 Stochastic models of evolution

The objective of a phylogenetic tree reconstruction method is to recover the *topology* of the evolutionary tree that gave rise to the observed taxa (typically represented by biomolecular sequences), since that topology indicates the order of speciation events that led to the observed taxa. However, for scientific reasons, the exact location of the root is often difficult to determine, so a method is considered to be accurate if it reconstructs the topology of the unrooted tree. In order to study different methods for evolutionary tree reconstruction, a model of biomolecular sequence evolution is assumed, and performance studies then reflect how well each method reconstructs the topology of different *model trees*, where a model tree is simply a rooted tree in which each edge is equipped with a stochastic model of evolution. Most typically, it is assumed that the sites (positions within the sequences) evolve identically and independently, and that the tree has the "Markov" property, so that the evolutionary processes that occur below a particular node in the tree do not depend upon what happens outside that subtree.

As an example of a simple model of sequence evolution, the *Cavender-Farris* [9] (also called "Cavender-Felsenstein") model is designed to describe the evolution of binary sequences (i.e. sequences of 0's and 1's). In this case, the evolutionary process is very simple. Every site evolves identically and independently, the root is drawn from some distribution (typically from the uniform distribution), and every edge e in the tree is associated with a probability $p(e)$, such that each site changes state on that edge with that probability. The $p(e)$ values need not be identical. When the sequences are over a larger alphabet (such as

for DNA, RNA, or amino-acid sequences), then for every edge e there is also an associated *mutation matrix* $M(e)$, which specifies the probability of changing between every pair of "states" (i.e. letters of the alphabet), given that a change occurs.

2.2 Phylogenetic tree reconstruction methods

Two types of methods generally are used to reconstruct trees from observed sequences. The first type is sequence-based, and uses the sequences directly to reconstruct the tree. Parsimony is probably the most popular sequence-based method, and it is based upon minimizing the number of mutations implied by the tree. This is computed by assigning sequences to every node in the tree, and then counting the number of changes that occur on each edge and adding these numbers up. Thus, parsimony is the "Hamming-distance Steiner tree problem". Maximum likelihood is the other most popular sequence-based method, and it seeks the tree which is most likely to have generated the observed sequences. The optimization problem, finding the most parsimonious tree, is known to be NP-hard even when restricted to binary characters [11, 20], although the optimal assignment of sequences to each internal node can be computed in polynomial time if the leaf-labelled tree is specified [19, 23]. However, in practice the heuristics used for parsimony seem to perform reasonably well, according to the folklore in systematic biology. By contrast, maximum likelihood is not known to be solvable in polynomial time if the input leaf-labelled tree is specified.

Distance-based methods, by contrast, operate by first computing distances between every pair of sequences, and then use these distances to reconstruct the tree. Thus, the input to a distance method is a matrix of observed distances, and the output is an edge-weighted tree whose distance matrix is close to the input matrix. There is an equivalent description of distance-based methods which is as follows. First, we define *additive* distance matrices to be matrices of leaf-to-leaf distances in edge-weighted trees in which all weights are positive. Since every edge-weighted tree defines an additive distance matrix, a distance-based method is actually a mapping from distance matrices to *additive* distance matrices, and as such we can study the properties of the different methods. One natural property we may wish to require is that a distance method *fix* additive matrices (we call this *combinatorial consistency*), and we may also wish to require that a distance method be *continuous*.

Many optimization problems related to distance-based reconstruction have been posed, but almost all have been shown to be NP-hard to solve exactly, and some are even hard to solve approximately [14, 1]. However, recently Agarwala *et al.* showed that a small modification to a classical method for tree reconstruction (proposed originally in [7, 8]) would provide a guaranteed performance ratio of 3 for the L_∞-nearest tree problem. This was later modified further by Cohen and Farach [9]. Both the Agarwala *et al.* and Cohen and Farach algorithms are 3-approximation algorithms, so that the output of these algorithms given distance matrix d is an additive distance matrix D such that $L_\infty(D, d) = max_{ij}|D_{ij} -$

$d_{ij}| \leq 3L_\infty(d, D^{opt})$, where D^{opt} is the nearest additive distance matrix to d under this metric.

2.3 Performance issues

Studies of the performance of different methods have been concerned generally with two distinct but related issues: *consistency*, i.e. whether the method *converges* to the correct topology as the sequence length increases, and *convergence rate*, which is the rate at which the error between the reconstructed tree and the true tree goes to 0, as the sequence length increases.

Performance studies (usually based upon simulations) on different model trees have revealed distinct differences between methods. For example, we now know (due to an analytically obtained result of Felsenstein [16], and confirmed experimentally in [27, 28]) that parsimony can make incorrect topology reconstructions, even if the sequences are unboundedly long (i.e. parsimony can be "inconsistent"), while almost all distance based methods will yield accurate reconstructions of the topology with arbitrarily high probability, if the sequences are long enough (i.e. distance methods are "consistent" throughout the parameter space). For some biologists, this has led to the rejection of parsimony as a method for reconstructing trees, since it can be "misled", while others have demonstrated that on some trees distance-methods may require sequence lengths that exceed that of genomes in order to be accurate [25].

2.4 Why distance methods are consistent

We provide a sketch of the proof of why all "reasonable" distance based methods are consistent, and note that full details can be obtained in [13].

Assume that T is a binary Cavender-Farris tree with n leaves. Thus, T is rooted, every non-leaf node has exactly two children, and we have associated to every edge e a mutation probability $p(e)$. Let sequences of length k evolve on this tree under the stochastic model of evolution implied by the tree, and define a distance on the sequences as follows:

$$d_{i,j} = -1/2ln(1 - 2H(i,j)/k),$$

where $H(i,j)$ is the hamming distance between i and j, i.e. $H(i,j)$ is the number of sites between sequences i, j in which they differ. Now let D be a distance matrix defined by $D_{ij} = lim_{k\to\infty} d_{ij}$. Then D is with probability 1 an *additive* metric (i.e. D is exactly equal to a matrix of path distances in an edge-weighted tree), and hence D defines a unique edge-weighted tree T'. Furthermore, the tree T' is actually the model tree, and the weights of the edges are exactly $w(e) = -1/2ln(1 - 2P(e))$.

It is worth noting that given any additive matrix D, the unique edge-weighted tree realizing D can be reconstructed in polynomial time (there are many such algorithms, the first of which is due to Waterman *et al.* [44]). Consequently, given D, not only can the topology of the model tree be obtained, but also

its mutation probabilities on the edges. This suggests a general approach to tree reconstruction based upon distances: *given distance matrix d, find a nearby additive metric D' and compute the edge-weighted tree corresponding to D'*. Since distance matrices that are computed from finite length sequences are close to the additive metrics that define the model tree that generated the sequences, such an approach will work if two following conditions can be met:

1. all "nearest" additive matrices define the topology of the model tree, and
2. a method is available which can recover a nearby additive matrix from an input matrix.

In [13], the conditions under which the first condition can hold were studied, and it was shown that there is a positive neighborhood (under the L_∞ metric) around every additive matrix corresponding to a binary edge-weighted tree such that every additive matrix in that neighborhood defined the same topology:

Theorem From [13]: *Let D be an additive distance matrix for an edge-weighted tree T which is binary and let $x > 0$ be the minimum weight of any edge in T. Then any additive distance matrix D' satisfying $L_\infty(D, D') < x/2$ defines a tree with the same topology as T.*

Consequently, we have the following:

Theorem From [13]: *If a distance method is* continuous *and* combinatorially consistent*, then the distance method is provably consistent for inferring binary trees.*

Proof. The proof is straightforward. Let T be an arbitrary Cavender-Farris model tree which is binary, and let x be the minimum weight of any edge under the transformation $w(e) = -1/2 \ln(1-2P(e))$. Let D be the additive matrix associated to T. Then let M be an arbitrary distance method which is continuous and combinatorially consistent. Since M is combinatorially consistent, $M(D) = D$. Since M is continuous, there is some $\epsilon > 0$ such that for all distance matrices d such that $L_\infty(d, D) < \epsilon$, $L_\infty(M(d), M(D)) < x/2$. But since $M(D) = D$, this implies that $M(d)$ and D have the same topology, which equals the topology of T. Finally, since the distances based on finite length sequences converge (with probability 1) on the distance matrix D, there will be some sequence length such that distances computed from sequences exceeding that length will be with high probability within ϵ of the distance matrix D, and in such a case the method will return the correct topology.

Almost all distance-based methods used in phylogeny estimation satisfy these two properties and hence are consistent methods for inferring binary Cavender-Farris trees. The only popular methods which violate these conditions are those that explicitly seek to reconstruct ultrametric trees (i.e. trees which are rooted so that the root is equidistant from all the leaves). Ultrametric tree reconstruction is appropriate when biomolecular sequences evolve at a more-or-less constant

rate; this hypothesis, otherwise known as the *molecular clock* hypothesis, has however been thoroughly discredited [6, 34, 31, 30, 43, 21], so that ultrametric tree reconstruction is no longer very much in favor.

2.5 Convergence rates of different methods

We have shown that almost all distance methods are consistent for inferring binary trees, but we have not discussed how well different methods perform at finite length sequences. However, the proof above provides a technique by which it is possible to infer something about the sequence length for which a method might be accurate. As indicated in the proof of Theorem 2, this inference takes two steps: In the first step we infer the largest ϵ so that $L_\infty(d, D) < \epsilon$ guarantees that $L_\infty(M(d), D) < x/2$, and in the second step we compute the sequence length needed to have $L_\infty(d, D) < \epsilon$ with high probability.

Using this two-step process, Erdös *et al.* proved the following:

Theorem From [13] *Let T is a Cavender-Farris model tree with edge mutation probabilities in the range $[f, g]$. Then for every $\epsilon > 0$ there is a constant c such that if sequences of length k are generated on this tree, where*

$$k > \frac{c \cdot \log n}{f^2(1-2g)^{2diam(T)}}, \tag{1}$$

then with probability $1 - \epsilon$ the result of applying the Neighbor-joining algorithm to corrected distances will be the true tree. The same formula (with the constant enlarged) exists for the Agarwala et al. *algorithm [1] and its variant, the Double-Pivot algorithm [10].*

The proof in [13] also showed that Neighbor-Joining could be guaranteed to be accurate from shorter sequences than the 3-approximation algorithms for the L_∞-nearest tree problem, and even from shorter sequences than an exact algorithm for the L_∞-nearest tree problem!

Note that the sequence length requirement for *guaranteed* accuracy goes up as either the lowest rate decreases, or as the highest rate increases. However, once we fix f and g, the lowest and highest mutation probability on any edge, then this theorem indicates a sequence length requirement that depends only upon the number n of leaves in the tree and the diameter of the tree. The diameter is the length (in terms of the number of edges) in the longest path in the tree. It is easy to see that the diameter always falls between $\log n$ and $n-1$, no matter what the tree, hence for every pair f, g, this sequence length requirement (for guaranteed accuracy) can vary dramatically from polylogarithmically to exponentially in n. Of importance, therefore, is the diameter of a typical tree that is studied in evolutionary tree reconstruction. However, the diameter of almost all trees is at least $\sqrt{n}$ [2], suggesting that the sequence length requirement for guaranteed accuracy generally will grow *superpolynomially* in the size of the tree. Or, in other words, neighbor-joining, the Agarwala *et al.* algorithm and its variant,

the Double-Pivot, might be incapable of inferring accurate topologies from large data sets.

However, this analysis is pessimistic; it implies good performance for these methods if the sequence length exceeds some lower bound, but does not correspondingly imply poor performance below that bound. Thus, the analytical result requires an experimental verification.

We studied this experimentally through a study [32] involving more than 200,000 data sets simulated on hundreds of model trees, and confirmed this conjectured bad performance. (A similar study was done by Strimmer and Von Haeseler, who found also that neighbor-joining based upon uncorrected distances required sequence lengths to grow exponentially in the number of taxa in the tree in order for a completely accurate topology estimation to be obtained.) However, we also saw that parsimony did not degrade in performance on large trees, at least not significantly, so that on large trees we observed that parsimony obtained accurate topology estimations from shorter sequences than neighbor-joining or the Agarwala *et al.* algorithm. Hillis also observed better performance by parsimony than by neighbor-joining in [24], although in that study the distinction between the methods was not as great as in ours.

Such results are potentially very disturbing, since molecular biologists have been relatively content with neighbor-joining because previous experimental studies tended to suggest that it had good performance [35]. However, despite intensive and extensive earlier experimental studies by a number of researchers (Felsenstein, Kuhner, Swofford, Huelsenbeck, and Hillis), this phenomenon had not been observed, because other studies had only rarely examined large enough trees, and had not systematically studied the effects of increasing size in the tree in order to make such an observation. Indeed, our own study in [32] shows that on small data sets, parsimony is in fact outperformed (albeit slightly) by neighbor-joining, and even by the Agarwala *et al.* algorithm [1], but that on larger trees the situation is definitely reversed, and the larger the tree, the greater the improvement of parsimony over distance-based methods.

However, parsimony is also problematic, since parsimony is known to be inconsistent in some portions of the parameter space [16, 27], and we do not know the conditions under which parsimony can be reliable (see [29] for a study which reveals that parsimony can be inconsistent even under conditions previously thought to be favorable to parsimony). Furthermore, parsimony is also NP-hard to solve exactly, and hence we do not have any reliable polynomial time algorithms to solve parsimony on all cases.

We summarize our observations as follows:

- parsimony is NP-hard to solve exactly, so that even a method which obtains optimal solutions may not reconstruct the correct topology everywhere (i.e. there are some trees on which parsimony will reconstruct the wrong topology with probability 1, even from unbounded length sequences),
- neighbor-joining uses only polynomial time, but it may not have adequate statistical power on large or highly divergent trees to reconstruct a reasonable estimate of the topology from realistic length sequences, and

– maximum likelihood estimation has great statistical power, but it is not generally used on large data sets due to its computational requirements.

The objective then is to develop methods which run in polynomial time and which obtain more accurate reconstructions of the model tree than existing methods, especially when reconstructing very large and divergent trees. We present in this paper a new method, which we call the *Short Quartet Method*, which seems to be capable of such a reconstruction. We describe its properties and present the results of an experimental study in which we compared the Short Quartet Method to Neighbor-Joining.

3 Short Quartet Method:

3.1 The method, briefly

We have developed a new quartet-based method called the *Short Quartet Method*. This method, originally introduced in [13], reconstructs trees based upon topologies of just a subset of the possible quartets containing the "short quartets". While the method used to reconstruct topologies on quartets can be arbitrary, Erdös *et al.* analyzed the performance of a variant of the method based upon using a simple distance-method to calculate topologies on quartets, as follows.

Definition 1. We denote by $ij|kl$ the topology on leaves i, j, k, l in which there is an internal edge separating i, j from k, l. The topology on i, j, k, l which has no internal edge separating two pairs of leaves is called the **star-topology**.

Relaxed four-point method:

Given sequences i, j, k, l, compute all distances between pairs in i, j, k, l. Return topology $ij|kl$ if $d_{ij} + d_{kl} < min(d_{ik} + d_{jl}, d_{il} + d_{jk})$. If all three pairwise sums are equal, return "star-topology".

The Short Quartet Method operates iteratively through a set of distance bounds, b. For each such bound b, it does the following:

– The set of quartets Q_b is computed, where $\{i, j, k, l\} \in Q_b$ if the maximum pairwise distance among i, j, k, l does not exceed b.

– The topology on every quartet in Q_b is computed using the relaxed four-point condition above, or some other method (for example, maximum likelihood or parsimony can be used).

– The topologies in Q_b are given as input to an algorithm called the *Short Quartet Consistency Algorithm* (defined in [13]). The output from the short quartet consistency algorithm is either a tree *uniquely* consistent with the topological constraints, or *failure*.

– If no unique tree is returned, then the bound b is increased. Eventually either the bound b exceeds the maximum distance observed in the input, or a tree is reconstructed which is uniquely consistent with the topologies.

3.2 Performance of the Short Quartet Method

As we have indicated, performance of evolutionary tree reconstruction methods is evaluated according to the degree of accuracy a method is likely to have on sequences of a given length generated on different model trees, and the specific different issues that are typically considered are *consistency* (i.e. accuracy from "long enough" sequences) and *convergence rate* (i.e. the rate at which the error goes to 0 as a function of the sequence length).

Erdös *et al.* studied the performance of the short quartet method, and proved that the method is consistent, so that under the assumption that the sequences evolve on a Cavender-Farris model tree (or, in fact, on the general Markov model!), once the sequences are long enough this greedy method would return the topology of the model tree with high probability. They also showed that the method runs in polynomial time since the short quartet consistency algorithm uses only polynomial time. The critical question then was how long the sequences had to be in order for this method to be accurate, or, to use the terminology we have introduced earlier, *how quickly does the short quartet method converge to the correct topology?*

Definition 2. If e is an edge in a tree T, then deleting the edge e (but not the endpoints of e) creates two rooted subtrees, t_1 and t_2. Let d_i be the distance from the root of t_i to its nearest leaf. Then the **depth** of the edge e is defined to be $max(d_1, d_2)$. The depth of T, denoted **depth(T)**, is the maximum depth of any edge in T. We say that i, j, k, l is a **short quartet** of a tree T if i, j, k, l are leaves in T and the maximum path length (counting only the number of edges) between any pair in i, j, k, l is at most $2depth(T) + 3$.

Erdös *et al.* made the following critical observations, upon which their method rests:

Theorem From [13]: *1. The short quartets suffice to define the tree, so that the set Q_b of quartet topologies defines the tree for all $b > B$, for some value B.*

2. If the Short Quartet Consistency algorithm is applied to Q_b for $b \geq B$, then it will reconstruct the unique tree consistent with the topologies in Q_b.

3. Hence, if all the quartets in Q_b are correct and $b \geq B$, then the topology of the model tree is returned.

It should be noted that the short quartet method either produces a tree consistent with all the quartet constraints, or it produces a star tree (i.e. the null tree). Thus, the short quartet method can fail entirely to obtain any information about the tree topology.

Erdös *et al.* studied the sequence length required for the Short Quartet Method to obtain a completely accurate reconstruction of the topology of the model tree.

Theorem From [13] *Let T is a Cavender-Farris model tree with edge mutation probabilities in the range $[f, g]$. Then for every $\epsilon > 0$ there is a constant c such that if sequences of length k are generated on this tree, where*

$$k > \frac{c \cdot \log n}{f^2(1-2g)^{4depth(T)}}, \tag{2}$$

then with probability $1 - \epsilon$ the result of applying the Short Quartet Method to corrected distances will be the true tree.

The comparison between the sequence length requirement of the Short Quartet Method to that of the sequence length requirement of the other methods (as provided by this analysis) thus depends on a comparison between the depth and the diameter.

The diameter of T (which we have denoted by $diam(T)$) is the number of edges in the longest path in T. It is clear that for small trees, $diam(T)$ is small as well, but for large trees $diam(T)$ can be quite large. The diameter of random trees has been analyzed in [2] and [13], showing that $diam(T)$ grows on the order of $\sqrt{n}$ for random trees under the uniform distribution [2], and on the order of $\log n$ for random trees under the Yule-Harding distribution [13]. These findings indicate that the sequence length needed for *guaranteed* accuracy by the Neighbor-Joining method, the Agarwala *et al.* [1] and Double-Pivota [10] 3-approximation algorithms for the L_∞-nearest tree, or *even* by an exact algorithm for the L_∞-nearest tree, can grow superpolynomially in the number of taxa in the dataset. On the other hand, the depth of a tree is never more than $\log n$, and in general can be shown to be bounded by $O(\log \log n)$ for random trees in either the uniform distribution or the Yule-Harding distributions [13]. Thus, the sequence length requirement for the Short Quartet Method generally grows *polylogarithmically*, and never more than polynomially, in the number of taxa in the dataset.

However, these results must then be evaluated experimentally, since they imply only that good performance is guaranteed in certain conditions, but do not imply bad performance when those conditions do not hold.

We present the results of our experimental performance analysis comparing neighbor-joining and the Short Quartet Method in Figure 1 and Figure 2. This study was obtained by simulating sequence evolution on a 50-taxon caterpillar in which we had uniform edge mutation probabilities on all edges except two most extreme edges, which each had mutation probabilities 5 times as great as the remaining edges. We varied the mutation probabilities on the edges while maintaining the ratio between every pair of edges, and we generated sequences of varying lengths. The horizontal axis represents sequence length, and the vertical axis represents overall sequence divergence. Grayscale values show fraction of 10 replicates for which the model tree was calculated correctly. The performance advantage enjoyed by the Short Quartet method over neighbor joining is very dramatic on this experimental study.

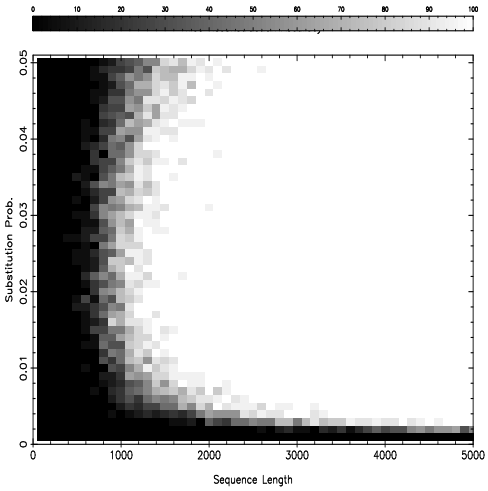

Fig. 1. Short Quartet Method on the 50-taxon caterpillar

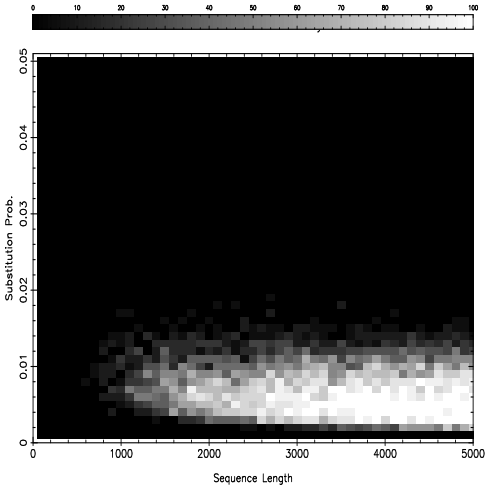

Fig. 2. Neighbor Joining on the 50-taxon caterpillar

4 Summary

The Short Quartet Method is a polynomial time method for tree reconstruction which operates by estimating topologies on quartets, and then combines these quartet topologies into one tree if they are consistent with each other and otherwise returns "failure". Thus, it is a quartet-based method in keeping with a long tradition; for example, the Berry-Gascuel method [5] for reconstructing the Buneman Tree, the Quartet Puzzling Method of Strimmer and Von Haeseler [37], the "network" construction method of Bandelt and Dress [4], and other classical

methods are all based upon estimating topologies on quartets and combining them. The way in which the Short Quartet Method is distinguished from previous quartet based methods is that it only uses a subset of the possible quartets, and the choice with which it selects those quartets allows an accurate topology estimation from much shorter sequences than in general is otherwise possible; i.e. the short quartet method has a faster convergence rate than other quartet-based methods. We have demonstrated this improvement in convergence rate through analytically obtained estimations of the sequence length needed for an accurate topology estimation, as well as through an experimental study, and have shown that this improvement can be quite dramatic.

It is important to note certain distinct aspects of this method. First, the method does not return a tree unless all the quartets topology estimations are consistent. This can mean that on some data sets there will be no tree returned, so that there is a disadvantage to using the short quartet method if the only objective is to obtain a tree. However, this can also be considered an advantage, since *if* a tree is obtained, there is an inherent *validation* of the tree that is significant and not generally obtained using other methods (which return trees even given completely random sequences).

Another distinct aspect of the method is that it does not seem to rely upon a completely accurate multiple sequence alignment. Rather, we never need sequence alignments of more than four sequences at a time, and the four sequences at a time are always close together in the tree and hence easier to align. Furthermore, we only need a sequence alignment which is accurate enough to indicate the correct topology of the quartet. This makes the problem of obtaining a multiple sequence alignment much easier than is usually the case for tree reconstruction, and avoids the problems of whether the tree reconstruction method is robust to errors in the multiple sequence alignment.

One of the major problems with reconstructing evolutionary trees which contain widely disparate taxa is the "missing data" problem; this occurs when, for example, there are genetic sequences that apply only to a small subset of the taxa. Most methods in such cases will require that only the portion of the sequences that apply to all the taxa be used for reconstruction purposes, thus resulting in shorter sequences rather than longer sequences for what is already a difficult reconstruction task. However, the short quartet method avoids this problem substantially. When some genetic sequences apply only to a subset of the taxa, these can still be used to reconstruct the subtrees on quartets that fall within that subset. This improves the accuracy of the topology prediction of the quartets, and hence of the entire tree.

The short quartet method also enables the use of different types of data, and does not specify the particular method by which each quartet is estimated (indeed, the choice of method for topology estimation can depend quite substantially on the particular quartet). Note also that both the analytical and experimental results were obtained for a distance-based variation of the short quartet method. In general, as extensive performance studies of different methods on four-taxon trees have shown [27], we would expect better accuracy from

shorter sequences using maximum likelihood or other more sensitive methods than the relaxed four-point method. We may also obtain greater accuracy by reconstructing subtrees on subsets larger than quartets, and we can still use the short quartet consistency algorithm to combine these subtrees when they are compatible. Thus, the short quartet method is a particular example of a general class of methods which compute trees on small subsets of taxa using an arbitrary method, and then combine the subtrees. This versatility with respect to subtree order and reconstruction method may make the method more robust to model violations.

5 Future Reading

Much interesting material can be obtained in the extensive literature in phylogenetics. We recommend in particular the following articles which provide an interesting survey of the field: [18, 39].

6 Acknowledgements

Tandy Warnow was supported by an NSF Young Investigator Award CCR-9457800, a David and Lucille Packard Foundation fellowship, and generous research support from the Penn Research Foundation and Paul Angello. Michael Steel was supported by the New Zealand Marsden Fund. Peter Erdös and Laszlo Szekely were supported in part by the Hungarian National Science Fund contracts T 016 358, T 019 367, and European Communities (Cooperation in Science and Technology with Central and Eastern European Countries) contract ERBCIPACT 930 113. Ken Rice was supported by postdoctoral fellowship in the University of Pennsylvania Computational Biology program, which is funded by NSF award BIR-9413215. This research started when the authors enjoyed the hospitality of DIMACS during the Special Year for Mathematical Support to Molecular Biology in 1995.

7 Bibliographic references

References

1. Agarwala, R., Bafna, V., Farach, M., Narayanan, B., Paterson, M. and M. Thorup. On the approximability of numerical taxonomy: fitting distances by tree metrics. *Proceedings of the 7th Annual ACM-SIAM Symposium on Discrete Algorithms,* 1996.
2. Aldous, D. J., Probability distributions on cladograms, in: *Discrete Random Structures*, eds. D. J. Aldous and R. Permantle, Springer-Verlag, IMA Vol. in Mathematics and its Applications. Vol. 76, 1-18, 1995.
3. Atteson, K. *Results on Neighbor-Joining's Convergence Rate*, to appear, Proceedings COCOON 1997.

4. Bandelt, H.-J., and A. Dress, Reconstructing the shape of a tree from observed dissimilarity data, *Adv. App. Math.*, **7**, 309–343 (1986).
5. Berry, V. and O. Gascuel, *Inferring evolutionary trees with strong combinatorial evidence*. To appear, proceedings of COCOON 1997.
6. Bruns, T.D. and T.M. Szaro, Mol Biol Evol 9 (5): 836-855 (Sep 1992) Rate and mode differences between nuclear and mitochondrial small-subunit rRNA genes in mushrooms.
7. Carroll, J.D. (1976). Spatial, non-spatial, and hybrid models for scaling, *Psychometrika*, 41 (4) 439-463.
8. Carroll, J.D., and S. Pruzansky. (1980) Discrete and hybrid scaling models. In *Similarity and Choice*, E.B. Lanterman and H. Freger, eds. Hans Huber, Berne.
9. Cavender, J. Taxonomy with confidence, *Mathematical Biosciences*, 40:271-280, 1978.
10. Cohen, J. and M. Farach. Numerical taxonomy on data: experimental results. Proceedings of the 1997 SODA.
11. Day, W.H.E., and D.S. Johnson, The computational complexity of inferring rooted phylogenies by parsimony. *Mathematical Biosciences*, 81:33-42, 1986.
12. Erdös, P., Steel, M. Szekely, L. and T. Warnow. 1997. Local quartet splits of a binary tree imply all quartet splits via one dyadic inference rule. To appear, Computers and Artificial Intelligence, special issue on Algorithms for Future Technologies.
13. Erdös, P., Steel, M., Szekeley, L. and T. Warnow. 1997. Inferring big trees from short sequences. Proceedings of International Congress on Automata, Languages, and Programming 1997.
14. Farach,M. Kannan, S. and T. Warnow. 1996. A Robust Model for Finding Optimal Evolutionary Trees. Algorithmica, special issue on Computational Biology, Vol. 13, No. 1, pp. 155-179. (A preliminary version of this paper appeared at STOC 1993.)
15. Farach, M. and S. Kannan, Efficient algorithms for inverting evolution, *Proceedings of the ACM Symposium on the Foundations of Computer Science*, 230–236, (1996).
16. Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading, *Syst. Zool.* **27** (1978), 401–410.
17. Felsenstein, J. PHYLIP – Phylogeny Inference Package (Version 3.2), *Cladistics,* 5: 164-166, 1989.
18. Felsenstein, J. Numerical methods for inferring evolutionary trees. *Quarterly Review of Biology*, 57 (1982), pp. 379-404.
19. Fitch, W.M. Towards defining the course of evolution. Minimum change for specific tree topology, *Syst. Zool.* **20** (1971), 406–416.
20. Foulds, L., and R. Graham, The Steiner problem in phylogeny is NP-complete, *Adv. Appl. Math.* **3**(1982), 43–49.
21. Gillespie, J.H. Proc. Natl. Acad. Sci. U S A 81 (24): 8009-8013 (Dec 1984) The molecular clock may be an episodic clock.
22. Green Plant Phylogeny Research Coordination Group, Summary report of Workshop #1: Current Status of the Phylogeny of the Charophyte Green Algae and the Embryophytes. University and Jepson Herbaria, University of California, Berkeley, June 24-28, 1995. 7 January, 1996.
23. Hartigan, J.A. Minimum mutation fits to a given tree, *Biometrics* **29** (1973), 53–65.
24. Hillis, D. Inferring complex phylogenies, *Nature* Vol **383** 12 September, 1996, 130–131.

25. Hillis, D. Huelsenbeck, J. and D. Swofford, Hobgoblin of phylogenetics? Nature, Vol. 369, 1994, pp. 363-364.
26. Hillis, D., Huelsenbeck, J., and C. Cunningham. 1994. Application and accuracy of molecular phylogenies. *Science*, 264:671-677.
27. Huelsenbeck, J.1995. Performance of phylogenetic methods in simulation. *Syst. Biol.* 44:17-48.
28. Huelsenbeck, J. and D. M. Hillis, Success of Phylogenetic Methods in the Four-taxon Case, *Syst Biol.*, 42:3 247-264, 1993.
29. Kim, J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. *Syst. Biol.* 45(3): 363-374.
30. Li, W.H. and M. Tanimura, Nature 326 (6108): 93-96 (Mar 5 1987) The molecular clock runs more slowly in man than in apes and monkeys.
31. Li, W.H., Tanimura M., and P.M. Sharp, J Mol Evol 25 (4): 330-342 (1987) An evaluation of the molecular clock hypothesis using mammalian DNA sequences.
32. Rice, K.A. and T. Warnow 1997. Parsimony is hard to beat! *COCOON97 Conference Proceedings*, in press
33. Saitou, N., and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406-425.
34. Sharp, P. and W.H. Li, J Mol Evol 28 (5): 398-402 (May 1989) On the rate of DNA sequence evolution in Drosophila.
35. Sourdis, J. and M. Nei, Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. *Mol. Biol. Evol.* (1996) 5:3 293-311.
36. Sogin, M.L., Hinkle, G. and D. D. Leipe, Universal tree of life, *Nature,* 362: 795, 1993.
37. Strimmer, K. and A. von Haeseler, *Quartet Puzzling: a quartet maximum likelihood method for reconstructing tree topologies,* Mol. Biol. Evol., 1996, 964–969.
38. Strimmer, K. and A. von Haeseler. 1996. Accuracy of Neighbor Joinging for n-Taxon Trees. *Syst. Biol.*, 45(4):516-523.
39. Swofford, D.L., Olsen, G.J., Waddell, P., and D. M. Hillis, Chapter 11: Phylogenetic inference, in: *Molecular Systematics*, D. M. Hillis, C. Moritz, B. K. Mable, eds., 2nd edition, Sinauer Associates, Inc., Sunderland, 1996, 407–514.
40. Swofford, D.L. PAUP: Phylogenetic analysis using parsimony, version 3.0s. Illinois Natural History Survey, Champaign. 1992.
41. Templeton, A. Human origins and analysis of mitochondrial DNA sequences. *Science* , Vol. 255, 737-739, 1992.
42. Wilson, A. C. and R. L. Cann, The recent African genesis of humans, *Scientific American* April 1992, 68–73.
43. Vawter, L, and Wm. Brown, Science 234 (4773): 194-196 (Oct 10 1986) Nuclear and mitochondrial DNA comparisons reveal extreme rate variation in the molecular clock.
44. Waterman, M.S., Smith, T.F., Singh, M., and W.A. Beyer, Additive evolutionary trees. *J. Theoret. Biol.*, 64:199-213, 1977.