Constructing Big Trees from Short Sequences

Péter L. Erdős
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] Biomathematics Research Centre, University of Canterbury. E-mail: m.steel@math.canterbury.ac.nz
[3] Department of Mathematics, University of South Carolina. E-mail: laszlo@math.sc.edu
[4] Department of Computer and Information Science, University of Pennsylvania. E-mail: tandy@central.cis.upenn.edu.

Abstract. The construction of evolutionary trees is a fundamental problem in biology, and yet methods for reconstructing evolutionary trees are not reliable when it comes to inferring accurate topologies of large divergent evolutionary trees from realistic length sequences. We address this problem and present a new polynomial time algorithm for reconstructing evolutionary trees called the *Short Quartets Method* which is consistent and which has greater statistical power than other polynomial time methods, such as Neighbor-Joining and the 3-approximation algorithm by Agarwala *et al.* (and the "Double Pivot" variant of the Agarwala *et al.* algorithm by Cohen and Farach) for the L_∞-nearest tree problem. Our study indicates that our method will produce the correct topology from shorter sequences than can be guaranteed using these other methods.

1 Introduction

Evolutionary trees indicate how species evolved from a common ancestor and are of fundamental concern to biologists. There are many methods for reconstructing trees from biomolecular sequences, and all potentially competitive methods are evaluated according to their accuracy for topology prediction [11]. However, reconstructing this topology is a difficult task for at least two reasons. First, all accepted optimization problems in this area are NP-hard, so that methods which are efficient typically do not provide good performance on large sets of sequences. More importantly, even if we could solve some of the NP-hard optimization problems in this domain, the sequence length required in order to be able to guarantee an accurate topology estimation can be beyond what is available or even possible. A polynomial time algorithm that can only be guaranteed to be accurate on unavailable sequence lengths is simply not reliable, and it must either not be used, or if used its output must not be believed. On the other hand, a method which is accurate on realistic length sequences *can* be used *even if* it requires more computational resources. We may simply need to use more machines, wait longer, employ more sophisticated techniques to implement the same basic objective, etc. Thus, the sequence length needed by a method imposes a significantly more severe limitation than its computational requirements. The importance to biologists of this measure of accuracy (called *efficiency* or *power* in the systematic biology literature [14]) is reflected in the extensive performance analysis literature in systematic biology in which methods are analyzed according to their performance on model tree reconstruction under various stochastic models of evolution [12]. Initially these studies focused on *consistency* [7], i.e. the question of whether a method would be guaranteed to produce the correct topology given long enough sequences. Since the discovery around 1970 [13] of *consistent distance transformations* (which produce *"corrected distances"*), it has been clear that all reasonable distance-based methods can recover the true tree with high probability given long enough sequences when applied to corrected distances computed on sequences generated by binary trees. All this is well-understood in the systematic biology community. What is not so well-understood is the sequence length needed to obtain an accurate topology with high probability using a given method on a given model tree. Unfortunately, sequence lengths are limited, and especially so when the tree to be reconstructed is large and contains widely divergent sequences.

This paper contains several results:

- We present a probabilistic analysis of the *depth* and *diameter* of random trees under two distributions.
- We describe a framework based upon *topology-invariant neighborhoods* which permits the comparison of the statistical power of different distance-based tree reconstruction methods.
- We develop a new consistent polynomial time method, the *Short Quartet Method* for reconstructing evolutionary trees, and provide an analytical study of its convergence rate for inferring trees under the Cavender-Farris model. (This analysis extends to a large class of r-state Markov models.) We show that this method has superior statistical power to Neighbor-Joining, the most popular distance-based method of phylogenetic tree reconstruction, and to new results from the theoretical computer science community by Agarwala *et al.* (STOC 1996) [1] and Cohen and Farach (SODA 1997 and RECOMB 1997) [5].

Due to space constraints, we cannot give proofs in this extended abstract.

2 Basics

We begin by describing a simple model of sequence evolution, called the *Cavender-Felsenstein* model, or sometimes the *Cavender-Farris* model. The Cavender-Felsenstein model of evolution for binary sequences associates to every edge e in a model tree T a *mutation probability* p_e with $0 < p_e < .5$, and the mutations on each edge are independent. The sites (i.e. positions within the sequences) are assumed to evolve identically and independently, with the state at the root selected according to some distribution (usually uniform). If k sites evolve under this model, then the tree generates a set of sequences of length k at the leaves. We allow the input to our method to be any symmetric zero-diagonal non-negative matrix, and we will abuse the notation and call such matrices *distance matrices*.

Definition 1. A distance matrix D is *additive* if and only if there exists a tree T with non-negative edge weighting w such that for all leaves i, j, $D_{ij} = \sum_{e \in P_{ij}} w(e)$, where P_{ij} is the path between i and j in T. The $\mathbf{L}_\infty$ distance between two distance matrices A and B is defined by $L_\infty(A, B) = max_{ij}|A_{ij} - B_{ij}|$. The $\mathbf{L}_\infty$-*nearest tree* problem takes as input a distance matrix d and returns an additive distance matrix D minimizing $L_\infty(d, D)$. The δ-*neighborhood* around d, denoted $N(d, \delta)$, is the set of all distance matrices d' such that $L_\infty(d, d') < \delta$. A *distance-based method* M for phylogeny construction is a mapping from $n \times n$ distance matrices to $n \times n$ additive distance matrices. A tree T_1 is said to *refine* a tree T if T can be obtained from T_1 by contracting some of the edges in T_1. A method M is said to be *combinatorially consistent* if $M(D) = D$ for all additive distance matrices D, and *continuous at* D if for every $\epsilon > 0$ there exists a $\delta > 0$ such that if $d \in N(D, \delta)$ then $M(d) \in N(M(D), \epsilon)$. We will say that a distance-based method is *reasonable* if it is both combinatorially consistent and continuous at every additive distance matrix defining a binary tree.

An interesting characterization of additive matrices D is the following:

Theorem 2. *Four Point Condition, from [4]: A distance matrix D is an additive matrix if and only if for all i, j, k, l, of the three pairwise sums $D_{ij} + D_{kl}, D_{ik} + D_{jl}, D_{il} + D_{jk}$, the largest two are identical.*

The proof of the theorem shows that the ordering on the three pairwise sums indicates the topology induced by the quartet. Thus, if $D_{ij} + D_{kl}$ is strictly smaller than the other two sums, then the topology induced by the quartet i, j, k, l is a resolved binary tree; otherwise all three sums are identical, and the topology induced by i, j, k, l is a star. Since we assume that T is binary, all such quartets induce resolved subtrees. We will denote this topology by $ij|kl$ when the pairs that are separated by an internal edge are ij and kl.

We now present a characterization of additive distance matrices which define the same topology.

Theorem 3. *Two additive distance matrices D and D' define the same topology if and only if for all quartets, the relative orders of the pairwise sums for that quartet are identical in the two matrices. Therefore, for every* reasonable *distance-based method M and for every binary tree T defining additive distance matrix D, there will be a $\delta > 0$ such that M is guaranteed to reconstruct the topology of T when applied to any $d \in N(D, \delta)$. Consequently, any* reasonable *distance-based method M will be* consistent *on every binary tree when applied to corrected distances. However, for every edge-weighted tree T with minimum* *edge weight x, there is a tree T' with a different leaf-labelled topology such that $L_\infty(D, D') = x/2$, where D is the additive distance matrix for T and D' the additive distance matrix for T'.*

We will now describe a method we call the *Naive Method*, based on Buneman's Four-Point Condition. For each quartet of species i, j, k, l, compute the topology on that quartet by computing the three pairwise sums (this is called the *four-point method* (FPM) for reconstructing a tree on a single quartet.) If the three sums are distinct and the minimum is attained at $D_{ij} + D_{kl}$, then set the topology on i, j, k, l to be $ij|kl$. If the minimum sum is not unique, constrain the topology to be a star. Construct the tree (if it exists) consistent with all the constraints on the topologies of quartets. If no tree exists consistent with all the constraints, output a star tree. (A similar procedure was described by Fitch in [9].) Constructing a tree consistent with all quartet topologies is easily done in polynomial time through a variety of techniques, hence this is a polynomial time method.

We now present a comparison of various distance based methods based upon topology invariant neighborhoods.

Theorem 4. *Let D be an additive $n \times n$ distance matrix defining a binary tree T, d be a fixed distance matrix, and let $\delta = L_\infty(d, D)$. Assume that x is the minimum weight of internal edges of T in the edge weighting corresponding to D.*
(i) *A hypothetical exact algorithm for the L_∞-nearest tree is guaranteed to return the topology of T from d if $\delta < x/4$.*
(ii) (a) *The 3-approximation algorithm for the L_∞-nearest tree is guaranteed to return the topology of T from d if $\delta < x/8$.* (b) *For all n there exists at least one d with $\delta = x/6$ for which the method can err.* (c) *If $\delta \geq x/4$, the algorithm can err for every such d.*
(iii) *The Naive Method is guaranteed to return the topology of T from d if $\delta < x/2$, and there exists a d for any $\delta > x/2$ for which the method can err.*

In other words, *given any matrix d of corrected distances, if an exact algorithm for the L_∞-nearest tree can be guaranteed to correctly reconstruct the topology of the model tree, then so can the Naive Method.* Thus, an exact algorithm for the L_∞-nearest tree *can err* on longer sequences than the Naive Method, when applied to corrected distances, for *any* model tree T. This suggests an inherent limitation of the L_∞-nearest tree approach to reconstructing evolutionary tree topologies.

3 The Short Quartet Method

The Short Quartet Method is similar in spirit to the Naive Method, in that it is based upon reconstructing trees for quartets, and then combining these trees if possible. However, the essential difference is that we attempt to avoid reconstructing the trees for the difficult quartets. Instead, we attempt to construct topologies only on those quartets that are close within the tree; these are called the *short quartets*. The reconstruction of the tree from these short quartets involves solving a special case of a problem which is in its general form NP-complete [15]. The method we use to reconstruct the topology on each quartet is not specified; if we can afford the time, we may elect to use maximum likelihood which has great statistical power, but which is computationally too expensive to use for all but small trees. However we do not know *apriori* which quartets are short quartets. Thus, the method we actually employ is a greedy method, which surprisingly can be shown to have high probability of accurate reconstruction of the topology provided that the sequence length is adequate, even if we reconstruct topologies on quartets using the same (simple and not particularly statistically powerful) method used by the Naive Method!

3.1 Short Quartet Consistency

We begin by defining the notion of an *edi*-subtree.

Definition 5. The *topological distance* between two leaves i and j in a tree T is the number of edges on the path between i and j, and the *topological length* of a path P is the number of edges on P. Consider the subtrees of a binary T obtained by deleting a single edge e in T but not the endpoints of e; call such subtrees *edi*-subtrees (for *edge-deletion-induced*). Each such edi-subtree can be considered a rooted tree, by rooting it at the endpoint of e to which it was originally attached. Given an *edi*-subtree t, $\mathbf{rep(t)}$ denotes a leaf in t closest to the root of t. Two *edi*-subtrees which are disjoint and whose roots are distance 2 apart are said to be *sibling edi*-subtrees. In order to simplify the discussion, we may abuse the notation and let t also denote the leaf set of the *edi*-subtree t.

We give some more definitions.

Definition 6. Let the *depth* of an edi-subtree in T be the number of edges on the path from e to the nearest leaf, and let the *depth* of T (denoted by $\mathbf{d(T)}$) be the maximum depth of any edi-subtree in T. We say that a path P in the tree T is *short* if its length is at most $2d(T)+2$. The quartet i, j, k, l is said to be a *short quartet* if it induces a subtree which contains a single edge connected to four disjoint *short* paths.

Thus, the depth of a complete binary tree of n leaves is $\log_2 n - 1$ but the depth of a caterpillar (a tree consisting of a long path with leaves hanging off the path) is just 1. Consequently, *every* quartet in a complete binary tree on n leaves is a short quartet, but there are only $O(n)$ short quartets in a caterpillar.

We now proceed with the description of the algorithm which we will use to construct binary model trees from a set of topologies on quartets. Our algorithm operates by determining siblinghood, first of leaves, and then of larger and larger rooted *edi*-subtrees, until the tree is constructed from the leaves inward. The determination of siblinghood of *edi*-subtrees is based upon detecting witnesses and anti-witnesses among the quartets, which we now define.

Definition 7. Given a quartet $\{i, j, k, l\}$ of leaves, we will denote by $ij|kl$ the induced topology on i, j, k, l in which i and j are separated in T from k and l via a path. Let t_1 and t_2 be two *edi*-subtrees. A *witness to the siblinghood of* t_1 *and* t_2 is a short quartet $\{u, v, w, x\}$ with topology $uv|wx$ such that $u \in t_1$, $v \in t_2$, and $\{w, x\} \cap (t_1 \cup t_2) = \emptyset$. We call such quartets *witnesses.* An *anti-witness to the siblinghood of* t_1 *and* t_2 is a short quartet $\{p, q, r, s\}$ with topology $pq|rs$, such that $p \in t_1$, $r \in t_2$, and $\{q, s\} \cap (t_1 \cup t_2) = \emptyset$. We will call these *anti-witnesses.*

We now present the property upon which the algorithm is based:

Axiom 1 *Let* t_1 *and* t_2 *be disjoint edi-subtrees of* T *and assume* $T - t_1 - t_2$ *has at least two leaves. Then* t_1 *and* t_2 *are siblings if and only if the following two conditions hold:*

1. *There are leaves* y *and* z *such that the quartet* $\{rep(t_1), rep(t_2), y, z\}$ *is a witness to the siblinghood of* t_1 *and* t_2*, and*
2. *If there is an antiwitness to the siblinghood of* t_1 *and* t_2*, then there is a witness for it as well.*

This axiom provides the basis for determining if there is at least one tree consistent with the constraints in the set of quartets, but may not be enough to verify that there are not two such trees. Verifying uniqueness of the solution turns out to be easy, fortunately, but it is also necessary due to the way in which we selectively apply the short quartet consistency algorithm.

In each *edi*-subtree, there may be more than one leaf that is closest to the root of the subtree (in terms of the number of edges on the path from the leaf to the root). However, among all such closest leaves in each *edi*-subtree, there is a unique leaf which has a smallest label, if the species are labelled by $1, 2, ..., n$. We call this leaf the **smallest representative** of the edi-subtree. This allows us to define a special set of short quartets, which we call the **representative quartets**, as follows. Each short quartet is composed of a single edge $e = (a, b)$, so that if we delete both a and b from T we create four *edi*-subtrees. We will say that a short quartet is a **representative quartet** if its leaves are the smallest representatives of the four *edi*-subtrees created in this manner. Then the following can be shown:

Theorem 8. *If a binary tree* T *is consistent with a set* Q *of quartet topologies such that* Q *contains all representative quartets, then* T *is uniqely consistent with* Q.

This observation and the axiom above suggests the following algorithm:

- Start with every leaf of T (i.e. the taxa) defining an *edi*-subtree.
- While the graph has more than three *edi*-subtrees, do:
  - Form the graph on vertex set given by the edi-subtrees, and with edge set defined by siblinghood; i.e., (x, y) is an edge if and only if edi-subtrees x and y satisfy the conditions of Axiom 1 for siblinghood.

    * Make a sibling pair out of each connected component, and make the roots of the *edi*-subtrees in that connected component children of a common root r, and replace the pair of *edi*-subtrees by one *edi*-subtree.
    * If no new sibling pairs are found, then return *fail*.
  - If there are at most three *edi*-subtrees left, connect their roots each to one internal node, and call the resultant tree T.
- Verify that T satisfies all the constraints given in the input, and that Q contains the *representative* quartet for every edge in T. If so, return T, and else return *fail*.

The correctness of this algorithm follows from the discussion above, and the runtime of this algorithm depends upon how the two *edi*-subtrees are found that can be siblings. It is obvious that this can be achieved in polynomial time, but the details of the implementation are omitted due to space constraints.

Theorem 9. *Given a set Q containing all short quartets of a tree T and satisfying Axiom 1, we can determine T in $O(|Q| \log n + n^2 \log n)$ time.*

3.2 The entire method

We now describe how we use the short quartet consistency algorithm to construct the tree. One issue we address is how we select the set of quartets to consider. As it turns out, this is done in a greedy fashion, which we now describe:

Definition 10. We define the **similarity** between sequences i and j to be $s(i,j) = 1 - 2H(i,j)/k$, where k is the sequence length, and $H(i,j)$ is the Hamming distance of sequences i and j. Let Q be the set of all possible quartets on $[n]$, and let Q_w be those quartets a, b, c, d such that $\min\{s(a,b), s(a,c), s(a,d), s(b,c), s(b,d), s(c,d)\} \geq w$.

On a given set Q_w, the result of applying the Short Quartet Consistency algorithm will either be a binary tree that is uniquely consistent with all the topology constraints in Q_w, or *fail*. This permits us to define our method as follows. The structure of the method is to do a *"halving"* search among the w by applying the Short Quartet Consistency algorithm to Q_w. starting with $w = 1/2, 1/4$, etc., until we either find a tree that is uniquely consistent with the Short Quartet consistency algorithm or realize that no such tree can be found (this evidence of failure occurs when $w < 1/k$). We can show that with high probability, given adequate sequence length this search will examine a set Q_w which contains all short quartets and which also satisfies Axiom 1. Consequently, in polynomial time we will reconstruct the tree topology.

Theorem 11. *The Short Quartets Method takes $O(n^4 \log n \log k + n^2 k)$ time in the worst case. On any input d of distances derived from sequences generated on a model tree T, if the Naive Method accurately reconstructs the topology of T from d then SQM will also accurately reconstruct the topology of T from d.*

A more realistic analysis of the running time of the Short Quartet Method is based upon analyzing *typical* trees can be obtained by using Theorem 13. Typical trees under both the uniform and Yule-Harding distributions have $O(\log\log n)$ depths. If the p_e probabilities on the edges of a tree of depth $O(\log\log n)$ are equal or almost equal, then certain Q_w's with $|Q_w| = O(n\ polylogn)$ will yield a tree through the consistency algorithm, and the halving search will hit such a w, with probability $1-o(1)$. Consequently, for typical tree shapes and for mutation probabilities that just slightly vary, applying the Short Quartet Method is likely to take only $O(n^2k + n^2\log n)$ time.

We now state our main result:

Theorem 12. *Suppose k sites evolve under the Cavender-Farris model on a binary tree T, so that for all edges e, $p_e \in [f, g]$, where we allow f, g to be functions of n. Assume that g is separated from $1/2$. The Short Quartet Method returns the tree T with probability $1-o(1)$, if*

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

where c is a fixed constant.

4 Depth vs. Diameter of Random Trees

We have shown that the sequence length needed by our method depends exponentially upon the minimum of the depth or the diameter of the tree it attempts to reconstruct. We study these topological quantities in this section.

Two simple models for describing semi-labelled binary trees are the *uniform* model, in which each tree has the same probability, and the *Yule-Harding* model, studied in [2, 3, 10]. This distribution is based upon a simple model of speciation, and results in "bushier" trees than the uniform model.

The following results are needed to analyse the performance of phylogeny reconstruction algorithms on random binary trees. Recall the definitions of depth and diameter from Section 3.

Theorem 13. *a) For a random semilabelled binary tree T with n leaves under the uniform model, $d(T) \leq (2+o(1))\log_2\log_2(2n)$ with probability $1-o(1)$, and $diam(T) > \epsilon\sqrt{n}$ with probability $1-O(\epsilon^2)$.*

b) For a random semilabelled binary tree T with n leaves under the Yule-Harding distribution, $d(T) = O(\log\log n)$ and $diam(T) = \Theta(\log n)$, with probability $1-o(1)$

4.1 Analysis of the Short Quartet Method

In [6], Farach and Kannan proposed a method (FK) for reconstructing Cavender-Farris trees based upon applying the 3-approximation of Agarwala et al (discussed in Section 2) for the L_∞-nearest tree problem to corrected distances. They proved that the method converged quickly for the *variational distance* (a related but different concern than the topology estimation), but did not analyze the convergence to the topology of the model tree. Recently, Kannan extended the analysis (personal communication) and obtained the following counterpart to (1): If T is a model tree with mutation probabilities in the range $[f, g]$, and if sequences of length k' are generated on this tree, where

$$k' > \frac{c' \cdot \log n}{f^2(1-2g)^{2diam(T)}}, \tag{2}$$

and c' is some constant, then with high probability the result of applying Agarwala et al to Cavender-Farris distances will be a tree with the same topology as T.

We now compare the sequence length requirements for the Short Quartet method as compared to the 3-approximation algorithm for the nearest L_∞-tree. Comparing this formula to (1), we note that the the comparison of depth and diameter is the most important issue. We always have $diam(T) \geq 2depth(T) + 1$. The constants do not affect the comparison unless the depth and the diameter are close to each other, which in general they are not (from our earlier results, for almost all trees, the depth is $O(\log\log n)$ while the diameter is $\Omega(\sqrt{n})$, under the uniform distribution, while for the Yule-Harding distribution, the depth is still $O(\log\log n)$ and the diameter is $\Omega(\log n)$. Consequently, the Short Quartet Method requires much shorter sequence lengths than the Agarwala et al algorithm for almost all binary trees.

We summarize these results in the following table.

		range of mutation probabilities on edges:	
		$[f, g]$ f, g are constants	$\left[\frac{1}{\log n}, \frac{\log\log n}{\log n}\right]$
binary trees	SQM	polynomial	polylog
worst-case	FK	superpolynomial	superpolynomial
random binary trees	SQM	polylog	polylog
(uniform model)	FK	superpolynomial	superpolynomial
random binary trees	SQM	polylog	polylog
(Yule-Harding)	FK	polynomial	polylog

This comparison establishes that our method requires significantly shorter sequences in order to ensure accuracy of the topology estimation than the algorithm of Agarwala et al, for almost all trees under both probability distributions. The trees for which the two methods need comparable length sequences are those in which the diameter and the depth are as close as possible – such as complete binary trees. In these cases, the previous analysis given in Section 3 indicates that SQM will nevertheless need shorter sequences than Agarwala et al will need to obtain the topology with high probability.

Although their running time is likely to be faster than ours on most data sets, our method is fast enough to be useful for all data sets that we might wish to analyze (even up to several thousand sequences). The real advantage of this method is its increase in accuracy on sequences of realistic length.

However, both algorithms are fast enough to make real-time computation of evolutionary trees feasible even for very large ($n = 500$ to 1000) data sets. This means that the issue of accuracy realistically is the most important issue, and needs to be the focus of the study.

5 Lower bounds

A careful analysis of the table above concerning the sequence length needed by the short quartet method reveals that for almost all trees under either distribution, the required sequence length grows polylogarithmically in the number of taxa for each fixed range of mutation probabilities. In this section, we show that this is a polynomial of the minimum possible sequence length for *any* method, whether deterministic or randomized.

We will henceforth assume that all trees we consider are binary trees bijectively leaf-labelled by the elements of $\{1, 2, \ldots, n\} = [n]$; we will call these *semi-labelled binary trees*. Since the number of semi-labelled binary trees on n leaves is $(2n-5)!!$, encoding deterministically all such trees by binary sequences at the leaves requires that the sequence length, k, satisfy $(2n-5)!! \leq 2^{nk}$, i.e. $k = \Omega(\log n)$. We now show that this information-theoretic argument can be extended for *arbitrary* models of evolution and *arbitrary* deterministic or even randomized algorithms for tree reconstruction. For each semi-labelled binary tree, T, and for each algorithm A, whether deterministic or randomized, we will assume that T is equipped with a mechanism for generating sequences, which allows the algorithm A to reconstruct the topology of the underlying tree T from the shortest possible sequences with constant probability.

Theorem 14. *Let T be a tree with n leaves labelled by sequences of $\{0,1\}^k$, and let A be an arbitrary algorithm, deterministic or randomized. For A to be able to reconstruct the topology of T from the sequences at the leaves with probability greater than 1/2 (respectively greater than ϵ), it must hold that $(2n-5)!! \leq 2^{nk}$ (respectively, $(2n-5)!!\epsilon \leq 2^{nk}$), and so $k = \Omega(\log n)$.*

The Theorem above shows that model and algorithm have to be a very good match, if not much more than $\log n$ length sequences suffice for tree reconstruction with high probability for each trees. In view of the very mild conditions, it is amazing, that this bound basically can be attained by our SQM, applied to the Cavender-Farris model!

6 Acknowledgements

Thanks to Ken Rice for carefully reading the manuscript for biological accuracy and Scott Nettles for advice about data structures. This research 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 to the fourth author. The second author was supported by the New Zealand Marsden Fund. The first and third authors 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. This research started when the authors enjoyed the hospitality of DIMACS during the Special Year for Mathematical Support to Molecular Biology in 1995.

References

1. R. Agarwala, V. Bafna, M. Farach, B. Narayanan, M. Paterson, 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. D. J. Aldous, 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. J. K. M. Brown, Probabilities of evolutionary trees, *Syst. Biol.* **43**(1), 78–91, (1994).
4. P. Buneman, The recovery of trees from measures of dissimilarity, in *Mathematics in the Archaeological and Historical Sciences*, F. R. Hodson, D. G. Kendall, P. Tautu, eds.; Edinburgh University Press, Edinburgh, 1971, pp. 387–395.
5. J. Cohen and M. Farach, Numerical Taxonomy on Data: Experimental Results. SODA '97 and RECOMB '97.
6. M. Farach, and S. Kannan, Efficient algorithms for inverting evolution, *Proceedings of the ACM Symposium on the Foundations of Computer Science*, 230–236, (1996).
7. J. Felsenstein, Cases in which parsimony or compatibility methods will be positively misleading, *Syst. Zool.*, **27**, 401–410 (1978).
8. J. Felsenstein, Numerical methods for inferring evolutionary trees, *Quarterly Review of Biology*, 57 (1982), pp. 379-404.
9. W. Fitch, A non-sequential method for constructing trees and hierarchical classifications. *J. Mol. Evol.*, (18):30-37, 1981.
10. E. F. Harding, The probabilities of rooted tree shapes generated by random bifurcation, *Adv. Appl. Probab.* **3**, 44–77, (1971).
11. D. Hillis, Approaches for assessing phylogenetic accuracy. *Syst. Biol.* **44**(1):3-16, 1995.
12. D. Hillis, J. Huelsenbeck, and D. Swofford, Hobgoblin of phylogenetics? Nature, Vol. 369, 1994, pp. 363-364.
13. J. Neyman, Molecular studies of evolution: a source of novel statistical problems. Pages 1-27 of Gupta, S.S. and J. Yackel (eds), *Statistical Decision Theory and Related Topics.* New York: Academic Press, 1971.
14. D. Penny, M. Hendy, and M. Steel, Progress with methods for constructing evolutionary trees. *Trends Ecol. Evol.* (7): 73-79, 1992.
15. M. A. Steel, The complexity of reconstructing trees from qualitative characters and subtrees, *J. Classification*, **9**, 91–116 (1992).