DIMACS Technical Report 96-19
July 1996

The number of nucleotide sites needed to accurately reconstruct large evolutionary trees[1]

by

Mike Steel
Biomathematics Research Centre, University of Canterbury
Christchurch, New Zealand

László A. Székely
Department of Computer Science, Eötvös University
1088 Budapest, Hungary

Péter L. Erdős
Department of Computer Science, University of Gödöllő
2103 Gödöllő, Hungary

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Research, Bellcore, and Bell Laboratories.

DIMACS is an NSF Science and Technology Center, funded under contract STC–91–19999; and also receives support from the New Jersey Commission on Science and Technology.

Abstract

Biologists seek to reconstruct evolutionary trees for increasing number of species, n, from aligned genetic sequences. How fast the sequence length N must grow, as a function of n, in order to accurately recover the underlying tree with probability $1-\epsilon$, if the sequences evolve according to simple stochastic models of nucleotide substitution? We show that for a certain model,a reconstruction method exists for which the sequence length N can grow surprisingly slowly with n (sublinearly for a wide range of parameters, and even as a power of $\log n$ in a narrow range, which roughly meets the lower bound from information theory). By contrast a more traditional technique (maximum compatibility) provably requires N to grow faster than linearly in n. Our approach is based on a new, and computationally efficient approach for reconstructing phylogenetic trees from aligned DNA sequences.

Keywords. Phylogenetic trees, DNA sequences, Azuma-Hoeffding inequality, phylogenetic invariants, subtrees.

[1]**Acknowledgment.** This research started when the authors enjoyed the hospitality of DIMACS during the Special Year for Mathematical Support to Molecular Biology. The first author gratefully acknowledges the New Zealand Ministry of Research, Science and Technology (MORST) for support to visit Budapest under ISAC Programme grant 94/22. Research of the second and third authors was supported in part by the Hungarian National Science Fund contract T 016 358 and by the European Communities (Cooperation in Science and Technology with Central and Eastern European Countries) contract ERBCIPACT 930 113.

1 INTRODUCTION

Simple models of nucleotide substitution are often used to analyse or justify methods for reconstructing evolutionary trees from aligned DNA sequences. One of the earliest, and still most striking examples of this approach, due to Felsenstein [14], reveals that two popular methods—*maximum parsimony* and *maximum compatibility*—can be seriously misled when the underlying mutation model has its parameters lying in a particular region (subsequently nicknamed "Felsenstein zone"). This result, and other more recent embellishments (see Hendy [17], Zharkikh and Li [28], Takezaki and Nei [26], Steel *et al.* [24]) concern statistical consistency, and as such are asymptotic results— that is they are concerned with outcomes as the sequence length (i.e. the number of sites) tends to infinity.

A more difficult question to analyse, particularly for large numbers of sequences, is how well methods perform for sequences of a given length, N. In particular, even if one is in a "good" region of the parameter space, it is clear that one needs at least a "reasonable" number of sites in order to be sure of recovering the correct tree by any method. Exactly how large this "reasonable" number must be, will surely depend also on the number of sequences (see Philippe and Douzery [20]), n.

More precisely, consider the question of how many sites N must be generated independently and identically, according to a substitution model M in order to reconstruct the underlying binary tree on n species with pre-specified probability $1-\epsilon$ by a particular method Φ. Clearly, the answer will depend on Φ, ϵ, and n, and also on the fine details of M—in particular the unknown values of its parameters. It is clear that for all models that have been proposed, if no restrictions are placed on the parameters associated with edges of the tree then the sequence length might need to be astronomically large, even for four sequences, since the "edge length" of the internal edge(s) of the tree can be made arbitrarily short (as was pointed out by Philippe and Douzery [20]). A similar problem arises for four sequences when one or more of the four non-internal edges is "long"—that is, when site saturation has occurred on the line of descent represented by the edge(s). Thus our question is interesting only if we assume that, for each tree, the parameters lie in some "good" region R (in the extreme, we might ask how long the sequences would need to be if the parameters were as favourable to us as possible).

This is related to, but different from, the question considered by Lecointre *et al.* [18]. In that paper the authors consider the length of sequences required in order for the reconstructed tree to be supported by high bootstrap proportions.

Before describing our results in more detail, we first provide a summary of notation used throughout this paper, and define more precisely the objects of study.

Notation: $\mathbb{P}[A]$ denotes the probability of event A; $\mathbb{E}[X]$ denotes the expectation of random variable X; all bold letters denote vectors, and if the coordinates of vector $\boldsymbol{x}$ are indexed by particular elements j we sometimes emphasis this by writing $\boldsymbol{x}= [x_j]$.

We denote the natural logarithm by log. The set $[n]$ denotes $\{1, 2, \dots, n\}$ and for any set S, $\binom{S}{k}$ denotes the collection of subsets of S of size k. $\mathcal{R}$ denotes the real numbers, and $\mathcal{R}_2[\boldsymbol{x}]$ denotes the vector space of quadratic polynomials in indeterminates $\boldsymbol{x} = [x_j]$ and coefficients in $\mathcal{R}$.

Definitions: (I) Trees. A *binary phylogenetic tree* (shortly *bph tree*) T is a tree whose *leaves* (vertices of degree 1) are labelled (by extant species, numbered $1, 2, \dots, n$) and whose remaining internal vertices (representing ancestral species) are unlabelled and of degree three. (Such trees are often assumed to represent the underlying evolutionary history of the collection of extant species.) Let $B(S)$ denote the set of bph trees on leaf set S, and let $B(n) = B([n])$. For $T \in B(n)$, $S \subseteq [n]$, there is a unique minimal subtree of T, containing all elements of S. We call this tree the *subtree* of T induced by S, and denote it by $T_{|S}$. We obtain the *binary subtree* induced by S, denoted by $T^*_{|S}$, if we substitute edges for all maximal paths of $T_{|S}$, in which every internal vertex has degree 2. Thus, $T^*_{|S} \in B(S)$. If $|S| = k$, then we refer to $T^*_{|S}$ as a *binary k-subtree.*

(II) Sites. Let us be given a set C of character states (such as $C = \{A, C, G, T\}$ for DNA sequences; $C =$ the 20 amino acids for protein sequences; $C = \{R, Y\}$ or $\{0, 1\}$ for purine-pyrimidine sequences). A sequence of length N is an ordered N-tuple from C—that is an element of C^N. A collection of n such sequences—one for each species labelled from $[n]$—is called a *collection of aligned sequences.* (In practice such a collection is derived from sequences of varying lengths by an "alignment" process, which aims to identify insertion and deletion events, and thereby to extract a subset of "sites" that differ between the sequences due to character substitutions). Aligned sequences have a convenient alternative description as follows. Let us call any map $\chi : [n] \to C$ a *pattern.* Then, a collection of n aligned sequences, each of length N, can equally well be represented as an ordered collection $\boldsymbol{s} = (s_1, \dots, s_N)$, of *sites*, where site s_i is the pattern that assigns $j \in [n]$ the character state at position i in sequence j. Let $\boldsymbol{x}[\boldsymbol{s}]$ be the vector, indexed by all possible $|C|^n$ patterns, and whose χ-coordinate is the proportion of sites where pattern χ occurs. Note that the map $\boldsymbol{s} \to \boldsymbol{x}[\boldsymbol{s}]$ represents $\boldsymbol{s}$ up to the order of the sites. In the rest of the paper we work with $C = \{0, 1\}$ only.

(III) Site substitution models. In this paper we assume that the sites are independently and identically distributed (i.i.d.) and are generated by some model, denoted M, which depends, in part, on the underlying bph tree, T_M. We let S_i denote the random variable site at position i, and let $\boldsymbol{S} = (S_1, ..., S_N)$; and define $\boldsymbol{x}[\boldsymbol{S}]$ as for $\boldsymbol{x}[\boldsymbol{s}]$ but with $\boldsymbol{s}$ replaced by the random variable $\boldsymbol{S}$. Let $f_\chi = \mathbb{P}[S_i = \chi]$, and $\boldsymbol{f}$, or more precisely, $\boldsymbol{f}(M) = [f_\chi]$. Thus, $\boldsymbol{f}$ equals the vector $\mathbb{E}[\boldsymbol{x}[\boldsymbol{S}]]$, and $\boldsymbol{x}[\boldsymbol{S}]$ has a multinomial distribution with parameters N and $\boldsymbol{f}$.

(IV) The Cavender–Farris model. Many models have been proposed to describe, stochastically, the evolution of sites. The simplest model, for two-state sites, is the symmetric model, due to Cavender [6] and Farris [13], which we have elsewhere called the CF (=Cavender–Farris) model. Let $\{0, 1\}$ denote the two states. Although the CF model is usually described, for biological reasons, on a rooted bph tree, we can, without loss of generality, disregard this feature of the model. For each edge e of T we have an associated *mutation probability*, which lies strictly between 0 and 0.5. Let $p : E(T) \rightarrow (0, 0.5)$ denote the associated map. Select one of the leaves, and assign it state 0 or state 1 with probability 0.5. Direct all edges away from this leaf and recursively assign random states to the vertices of T as follows: if $e = \{u, v\}$ is directed from u to v, and u (but not v) has a state assignment, then v is assigned the same state as u with probability $1 - p_e$ or the other state with probability p_e. It is assumed that all assignments are made independently, and so the pair (T, p) determines the joint probability of any assignment of states to the vertices of T, and thereby the marginal probability of any assignment of states to the leaves of T—and this then provides a probability distribution on all binary patterns on $[n]$. We let $\boldsymbol{f}$ denote the vector of these 2^n pattern probabilities.

The CF model is hereditary on subsets of the leaves—that is, if we select a subset S of $[n]$, and form the binary subtree $T^*_{|S}$, then we can define mutation probabilities on the edges of $T^*_{|S}$ so that the probability distribution on the patterns on S is the same as the marginal of the distribution on patterns provided by the orginal tree T. Furthermore, the mutation probabilities that we assign to an edge of $T^*_{|S}$ is just the probability p that the endpoints of the associated path in the original tree T are in different states, and p is nicely related to the mutation probabilities $p_1, p_2, ..., p_k$ of edges of the k-path of the original tree:

$$p = \frac{1}{2}\left(1 - \prod_{i=1}^{k}(1 - 2p_i)\right) . \tag{1}$$

Formula (1) is well-known, and it is easy to prove by induction.

(V) Tree reconstruction. A *phylogenetic tree reconstruction method* is a function Φ that associates to every collection of sites either a bph tree, or the statement `fail`, indicating that the method is unable to make such a selection for the data given. Many such methods have been proposed, and mostly these are invariant under permutation of the sites. Thus, we will regard Φ as operating on the vector $\boldsymbol{x}[\boldsymbol{s}]$, or more generally, on the $(K-1)$-dimensional simplex (where $K = |C|^n$) : $\{(x_1, ..., x_K) : x_i \geq 0, \sum x_i = 1\}$ in which $\boldsymbol{x}[\boldsymbol{s}]$ sits.

It is essential for tree reconstruction that two different bph trees cannot underlie models that produce the same distribution on sites—that is

$$T_M \neq T_{M'} \text{ implies } \boldsymbol{f}(M) \neq \boldsymbol{f}(M') . \tag{2}$$

A case where (2) is violated arises when there is an unknown distribution of rates across sites as described by Steel *et al.* [24]. However, provided the sites evolve i.i.d. under a suitable Markov-style assumption condition, (2) holds (Steel [23], Chang [8]). Now we discuss a specific, popular tree reconstruction method.

(VI) Maximum compatibility. Given a bph tree T with leaf set $[n]$, deleting an edge e of T disconnects T into two components, and thereby induces a bipartition of

$[n]$ consisting of the leaves of the two components. This bipartition is called a *split* of T induced by the edge e; the split is called *non-trivial*, if both components contain at least 2 leaves. Now each site also gives a bipartition of $[n]$ by grouping together the taxa that have the same character state at that site. If this bipartition is a non-trivial split of T, the site is said to *support* T, and in this case we also call the site *non-trivial*. The *maximum compatibility method* selects a tree that maximizes the number of supporting sites.

Buneman [4] showed that each bph tree T is uniquely defined by its non-trivial splits. This fact justifies the use of the maximum compatibility method.

Of fundamental interest in phylogenetic analysis is the probability of recovering the correct tree, that is $\mathbb{P}[\Phi(\boldsymbol{x}[\boldsymbol{S}]) = T_M]$. This probability is dependent on M, Φ and N, the number of sites. The method Φ is said to be *statistically consistent* for a class ϑ of models M, if for all $M \in \vartheta$, $\lim_{N\to\infty} \mathbb{P}[\Phi(\boldsymbol{x}[\boldsymbol{S}]) = T_M] = 1$. In this paper we are concerned with the question of how fast N must grow as function of n in order for $\mathbb{P}[\Phi(\boldsymbol{x}[\boldsymbol{S}]) = T_M]$ to remain close to 1.

In Section 2 we show that for the method of maximum compatibility, the number N of sites must grow faster than linearly in n, for this reconstruction probability to be at least $1 - \epsilon$ for ϵ fixed. This is regardless of the details of the model M, or the values any parameters in this model may take. This result is hardly surprising, but it makes a useful contrast to our main, and somewhat surprising result in Section 3.

The main result of the paper is a method (which is a polynomial time algorithm in nN) for reconstructing trees from sequences developed under the Cavender–Farris model. For this method N does not need grow very quickly with n—indeed slower than n—provided the underlying parameters in the model (related to the times between speciation events) lie in a certain range.

There is a simple information theoretic lower bound for N, if tree reconstruction is possible. The number of bph trees with n labelled leaves is $|B(n)| = (2n-5)!!$ [5]. If all bph trees are encoded with N sites, each site has 2^n character states, and all trees can be reconstructed (for sure, or for almost sure), then we must have $(2n-5)!! \le 2^{nN}$, i.e. $c \log n < N$. Surprisingly enough, setting the transition probabilities in the CF model in the proper narrow range, we get very close to this bound by our reconstruction method. As far as we are aware, these are the first such analytic results in this area.

It is important to stress that we are not advocating the use of the method we describe. For a small n it may well require a sequence length N larger than other more conventional statistical techniques, such as maximum likelihood (Felsenstein [15], Goldman [16], Saitou [21]). Furthermore the bounds we give for our method are also, almost certainly, not the best possible. Our results are described for a two-state model, but it seems likely that similar results apply for models on four (or an arbitrary number) or states.

The main result required the development of algorithmic techniques to reconstruct bph trees from "local" binary 4-subtrees (Theorem 2), a new method to reconstruct 4-leaf trees (Theorem 3), and probabilistic techniques to extend this to a proceedure on n-leaf trees. At the end of the paper we illustrate a further application to the study of quadratic invariants (Proposition 1).

2 MAXIMUM COMPATIBILITY: A LOWER BOUND ON N FOR ALL MODELS

We first show that a simple, conventional method—*maximum compatibility*—requires a superlinear sequence length in order to recover the correct tree with close-to-one probability, regardless of how favourable are the parameters in the underlying model.

Theorem 1 *Assume that sites on n species evolve according to any model M of nucleotide substitution (as in Section 1 definition (III)). Suppose the maximum compatibility method Φ_{MC} is applied to reconstruct T_M.*

If $N(n)$ denotes the smallest number of sites for which $\mathbb{P}[\Phi_{MC}(\boldsymbol{x}[\boldsymbol{S}]) = T_M] \geq 1/2$, then for n large enough,

$$N(n) > (n-3)\log(n-3) - (n-3). \tag{3}$$

Proof. Assume that we are given $N(n) \leq (n-3)\log(n-3) - (n-3)$ sites, and the the number of non-trivial sites among them is less or equal to N^*, the smallest integer greater or equal to $(n-3)\log(n-3) + x(n-3)$. We will show that the probability of obtaining the correct tree under Φ_{MC} is at most $e^{-e^{-x}}$, which proves the theorem by setting $x = -1$, since $N(n) \geq N^*$.

Let $\sigma(T)$ denote the set of non-trivial splits of $T = T_M$. We have $|\sigma(T)| = n-3$ [4]. For $\sigma \in \sigma(T)$, let the random variable X_σ be the number of non-trivial sites which induce split σ. Let $X := \sum\limits_{\sigma\in\sigma(T)} X_\sigma$. A necessary (though not sufficient) condition for maximum compatibility to select T is that all the non-trivial splits of T are present amongst the N^* non-trivial sites. Thus, we have the inequality:

$$\mathbb{P}[\Phi_{MC}(\boldsymbol{x}[\boldsymbol{S}]) \ = \ T_M] \leq \mathbb{P}[\cap_{\sigma\in\sigma(T)}\{X_\sigma > 0\}] \ .$$

Conditioning on the size of X,

$$\mathbb{P}[\cap_{\sigma\in\sigma(T)}\{X_\sigma > 0\}] \ = \ \sum_k \mathbb{P}[\cap_{\sigma\in\sigma(T)}\{X_\sigma > 0\} \mid X = k] \times \mathbb{P}[X = k]$$

$$\max_{1\leq k\leq N^*} \mathbb{P}[\cap_{\sigma\in\sigma(T)}\{X_\sigma > 0\} \mid X = k] \ =$$

$$\mathbb{P}[\cap_{\sigma\in\sigma(T)}\{X_\sigma > 0\} \mid X = N^*] \ . \tag{4}$$

Let $p(\sigma)$ denote the probability of generating split σ at a particular site. Due to the model, $p(\sigma)$ does not depend on the site. We will show that (4) is maximized when the $p(\sigma)$'s are all equal ($\sigma \in \sigma(T)$) and sum to 1. In that case, determining (4) is just the classical occupancy problem where N^* balls are randomly assigned to $(n-3)$ boxes with uniform distribution, and one asks for the probability that each box has at least one ball in it. Equation (3) now follows from a famous result concerning the asymptotics of this problem (Erdős and Rényi [12]): for $x \in \mathcal{R}$, N^* balls (N^* as defined above), and $(n-3)$ boxes, the limit of probability of filling each boxes is $e^{-e^{-x}}$.

¿From compactness arguments, there exists a probability distribution maximizing (4). We show that it cannot be non-uniform, and therefore the uniform distribution maximizes (4). Assume that the maximizing distribution p is non-uniform, say, $p(\sigma) \neq p(\rho)$. We introduce a new distribution p' with $p'(\sigma) = p'(\rho) = \frac{1}{2}(p(\sigma) + p(\rho))$, and $p'(\alpha) = p(\alpha)$ for $\alpha \neq \sigma, \rho$. The probability of having exactly i sites supporting σ *or* ρ is the same for p and p'. Conditioning on the number of sites supporting σ or ρ, it is easy to see that any distribution of sites supporting all non-trivial splits has strictly higher probability in p' than in p. □

3 AN UPPER BOUND ON N FOR THE CAVENDER–FARRIS MODEL

In this section we describe a tree reconstruction method for which the sequence length can (for certain models) grow relatively slowly as a function of the number of species, in order that the correct tree be recovered with high probability.

We first discuss how partial information on binary 4-subtrees of T can be used to determine T. Then we provide for a novel technique to reconstruct the binary 4-subtrees.

Finally we give an algorithm that uses the two techniques discussed above as procedures to reconstruct trees on n species that has the claimed sublinear performance when the parameters in the underlying model lie in a certain region.

3.1 RECONSTRUCTING A BPH TREE FROM BINARY 4-SUBTREES

For a bph tree $T \in B(n)$, and a quartet of leaves, $q \in \binom{[n]}{4}$, let $L_T(q)$ denote the length of the longest path of $T_{|q}$ which turned into an edge of $T^*_{|q}$. For $q = \{a, b, c, d\}$ we say that $t_q = ab|cd$ is a *valid quartet split* of T, if $ab|cd$ is a split of $T^*_{|q}$. As in Bandelt and

Dress [2], it is easy to see that

$$\text{if } ab|cd \text{ is a valid quartet split of } T, \text{ then so are } ba|cd \text{ and } cd|ab, \tag{5}$$

and we identify these three splits. If (5) holds, then $ac|bd$ and $ad|bc$ are not valid quartet splits of T, and we say that any of them *contradicts* to (5). Also,

$$\text{if } ab|cd \text{ and } ac|de \text{ are valid quartet splits of } T, \text{ then so are } ab|ce,\ ab|de, \text{ and } bc|de, \tag{6}$$

and,

$$\text{if } ab|cd \text{ and } ab|ce \text{ are valid quartet splits of } T, \text{ then so is } ab|de. \tag{7}$$

Let $Q(T) = \left\{t_q : q \in \binom{[n]}{4}\right\}$ denote the set of valid quartet splits of T. It is a classical result that $Q(T)$ determines T (Colonius and Schulze [10], Bandelt and Dress [2]); indeed for each $i \in [n]$, $\{t_q : i \in q\}$ determines T, and T can be computed in polynomial time. For example, a simple algorithm for reconstructing T from $Q(T)$ is simply to build up T recursively from the tree with leaf set 1,2,3 by attaching (in any order) the remaining elements from $[n]$ as new leaves to the tree so far constructed. In this way, one uses $Q(T)$ to determine the unique edge of each partial tree to which the new leaf must be attached by bisecting the edge and making the newly created vertex adjacent to the new leaf.

An extension of this result is that for any $T \in B(n)$ a carefully chosen subset of $Q(T)$ of cardinality $n-3$ determines T (Steel [22]). Another extension is that an unknown bph tree T with n leaves can be constructed by asking at most $O(n \log n)$ queries of the form: "what is t_q?" for a choice of q that depends on the answers to the queries so far asked (Pearl and Tarsi [19], Warnow [27]).

It would be useful to tell from a set of quartet splits if they are valid quartet splits of any bph tree. Unfortunately, this problem is NP-complete (Steel [22]). It also would be useful to know, which subsets of $Q(T)$ determine T and which subsets would allow for a polynomial time procedure to reconstruct T. A natural step in this direction is to define *inference*: a set of quartet splits A infers a quartet split t, if whenever $A \subseteq Q(T)$ for a bph tree T, then $t \in Q(T)$ as well.

Setting a complete list of inference rules seems hopeless (Bryant and Steel [3]). Instead, Dekker [11] introduced a restricted concept, *dyadic* and higher order inference. He says that a set of quartet splits A *dyadically infers* a quartet split t, if t can be derived from A by repeated applications of rules (5), (6) and (7). We say that a set of quartet splits A *semidyadically infers* a quartet split t, if t can be derived from A by repeated applications of rules (5), (6).

Quartet splits (semi)dyadically inferred by a set of quartet splits can be computed in polynomial time, and quartet splits (semi)dyadically inferred by a set of valid quartet splits of a tree are valid. We denote by $cl_2(A)$ the set of quartet splits semidyadically inferred by a set of quartet splits, A. We say that a set of quartet splits A *(semi)dyadically determine* T, if they (semi)dyadically infer *all* valid quartet splits of T, i.e. $Q(T)$.

Here we provide a third extension of Colonius and Schulze's classical result, by showing that for a binary tree T, the subset of $Q(T)$ consisting of those quartets q, for which $L_T(q) \leq c \log n$, determines T, where c is a constant. We show that c can be taken to be 18, which suffices for the proof of our main result (Theorem 4), though with more work this value can be reduced further (see comment (3) in Section 4).

Theorem 2 *For a bph tree T on $[n]$ ($n \geq 4$), let*

$$D(T) = \{q \in \binom{[n]}{4} : L_T(q) \leq 18 \log n\}.$$

Then $S(T) = \{t_q$ valid quartet split of T $: q \in D(T)\}$ semidyadically determines T. In particular, T can be reconstructed from $S(T)$ in polynomial time.

Proof. We use induction on n. Let us be given $T \in B(n)$. There is an edge e_1 in T, which defines a split of $[n]$ into classes as equally-sized as possible. Then each class has $\geq n/3$ leaves. (For if not, one can find a split even closer to equal by considering the split induced by the edge connecting e_1 to the bigger subtree on the big side.) Let $A \cup B$ and $C \cup D$ denote the classes of the split. Let $T_{|A\cup B}$ and $T_{|C\cup D}$ denote the two subtrees of T obtained by the deletion of e_1. A similar argument would provide for an edge e_2 of $T_{|A\cup B}$ and e_3 of $T_{|C\cup D}$, so that each side of the split of $T_{|A\cup B}$, say A and B, and each side of the split of $T_{|C\cup D}$, say C and D, has at least $n/9$ leaves. We make, however, an extra condition:

$$e_2 \text{ and } e_3 \text{ are vertex disjoint from } e_1, \tag{8}$$

thus we achieve only that each side of the split of $T_{|A\cup B}$ and $T_{|C\cup D}$ is at least $n/18$. This partitioning with e_1, e_2, and e_3 fails only if $T_{|A\cup B}$ or $T_{|C\cup D}$ has two leaves only. Then $n \leq 5$ or $n = 6$ and T has no path longer than four. These are the base cases for our induction, the quoted theorem of Colonius and Schulze yields the proof of the base cases, since $S(T) = Q(T)$ holds for them.

For the induction step, consider the leaf partition we just defined:

$$A \underset{}{\overset{e_2}{\rule{0pt}{0pt}\_\_\_\_\_}} B \overset{e_1}{\_\_\_\_\_} C \overset{e_3}{\_\_\_\_\_} D \tag{9}$$

Recall that $|B| \geq 2$ and $|C| \geq 2$ by (8). Let $T^*_{|A\cup B\cup C}$ denote the left binary subtree of $T - e_3$, and let $T^*_{|B\cup C\cup D}$ denote the right binary subtree of $T - e_2$. Observe that $T^*_{|A\cup B\cup C}$ and $T^*_{|B\cup C\cup D}$ each has at least 5, but at most $(17/18)n$ leaves.

Assume that $t = t_q \in Q(T)$. We have to show $t \in cl_2(S(T))$. We do it through investigating the distribution of the elements of q in A, B, C, D. We neglect giving refences to (5) in the proof.

Case left: $q \subset A \cup B \cup C$.
First we show that $S(T^*_{|A\cup B\cup C}) \subseteq S(T)$. Take any $q' \in S(T^*_{|A\cup B\cup C})$. Note that $18\log n \geq 18\log(17n/18)+1 \geq L_{T^*_{|A\cup B\cup C}}(q')+1 \geq L_T(q')$, thus $q' \in S(T)$. Since cl_2 is monotone, $t \in cl_2(S(T^*_{|A\cup B\cup C}))$ will imply $t \in cl_2(S(T))$. ¿From the drawing (9) $t = t_q \in Q(T^*_{|A\cup B\cup C})$ as well, and using the hypothesis $t \in cl_2(S(T^*_{|A\cup B\cup C}))$.

Case right: $q \subset B \cup C \cup D$.
Exchange $T^*_{|A\cup B\cup C}$ to $T^*_{|B\cup C\cup D}$ in the proof of the previous case.

In the rest of the proof lower case letters denoting leaves indicate as well the partition class where they belong to. Due to the first two cases settled above, any $t = t_q \in Q(T)$, for which we still have to show $t \in cl_2(S(T))$, has the property that q intersects A and D. Case xyuv below means that $|q \cap A| = x$, $|q \cap B| = y$, $|q \cap C| = u$, $|q \cap D| = v$. Using the left-right symmetry of the drawing we further reduce the number of cases. We neglect references in the proof to (5).

Case 1111: $ab|cd \in Q(T)$.
Let e_4 denote an edge which separates e_1 and e_3. Then we also have two leaves $c, c' \in C$, separated by e_4. Edge e_1 in drawing (9) shows that $ab|cc' \in Q(T)$. Edge e_4 shows that either $bc|c'd \in Q(T)$ or $bc'|cd \in Q(T)$, but $bd|cc' \notin Q(T)$. In the first case the sought for quartet is inferred by $ab|cc' \in cl_2(S(T))$ (Case left) and $bc|c'd \in cl_2(S(T))$ (Case right) by (6). In the second case use $bc'|cd \in cl_2(S(T))$ (Case right) instead of $bc|c'd \in cl_2(S(T))$ (Case right).

Case 2101: $aa'|bd \in Q(T)$.
By e_2, $aa'|bc \in Q(T)$ and by e_1, $a'b|cd \in Q(T)$. Hence $aa'|bc \in cl_2(S(T))$ (Case left) and $a'b|cd \in cl_2(S(T))$ (Case 1111); (6) finishes the proof.

Case 2011: $aa'|cd \in Q(T)$.
By e_2, $aa'|bc \in Q(T)$ and by e_1, $ab|cd \in Q(T)$. Hence $aa'|bc \in cl_2(S(T))$ (Case left) and $ab|cd \in cl_2(S(T))$ (Case 1111); (6) finishes the proof.

Case 2002: $aa'|dd' \in Q(T)$.
By e_1, $aa'|cd \in Q(T)$, and by e_3, $ac|dd' \in Q(T)$. Hence $aa'|cd \in cl_2(S(T))$ (Case 2011) and $ac|dd' \in cl_2(S(T))$ (symmetry to Case 2101); (6) finishes the proof.

Case 3001: $a_1a_2|a_3d \in Q(T)$.
Note that $a_1a_2|a_3b \in Q(T)$, from the drawing (9), and $a_1a_2|a_3b \in cl_2(S(T))$ (Case left). By e_2, $a_1a_3|bd \in Q(T)$, and $a_1a_3|bd \in cl_2(S(T))$ (Case 2101). Using (6) finishes the proof.

Case 1201: Subcase $ab_1|b_2d \in Q(T)$.
Note that $ab_1|b_2c \in Q(T)$, from the drawing (9), and by e_1, $b_1b_2|cd \in Q(T)$. We have $ab_1|b_2c \in cl_2(S(T))$ (Case left) and $b_1b_2|cd \in cl_2(S(T))$ (Case right). Using (6) finishes the proof.
Subcase $ad|b_1b_2 \in Q(T)$.

Note that $ac|b_1b_2 \in Q(T)$, due to the subcase that we are in and drawing (9). By edge e_1, we have $ab_1|cd \in Q(T)$. We have $ac|b_1b_2 \in cl_2(S(T))$ (Case left) and $ab_1|cd \in cl_2(S(T))$ (Case 1111). Using (6) finishes the proof.

We proved that $S(T)$ semidyadically determines $Q(T)$. Now there is an obvious polynomial time algorithm to reconstruct T: look for new quartet splits semidyadically inferred by $S(T)$, and when you have all $\binom{n}{4}$ quartet splits, use the Colonius–Schulze algorithm to reconstruct T. $\square$

3.2 RECONSTRUCTING BINARY SUBTREES ON FOUR SPECIES

There are numerous techniques for reconstructing trees for four species. In this section we construct a method Φ^1 for which we derive a useful lower bound on $\mathbb{P}[\Phi(\boldsymbol{x}[\boldsymbol{S}]) = T]$ for any sequence length N when sites evolve under the CF model. We start with some prerequisites.

For $N > 1$ consider the linear transformation ψ_N on $\mathcal{R}_2[\boldsymbol{x}]$ given by:

$$\psi_N \left[\sum_{i,j} c_{ij} x_i x_j + \sum_i d_i x_i + e \right] = \sum_{i,j} c^*_{ij} x_i x_j + \sum_i d^*_i x_i + e \tag{10}$$

where $c^*_{ij} = \frac{N}{N-1}\, c_{ij}$ and $d^*_i = d_i - \frac{1}{N-1}\, c_{ii}$.

The following two lemmas will be useful later, and are easily established.

Lemma 1 *Suppose $\boldsymbol{X} = [X_i]$ has a multinomial distribution with parameters $N > 1$ and $\boldsymbol{\pi} = [\pi_i]$. For any $p \in \mathcal{R}_2[\boldsymbol{x}]$, $\psi_N \left[p\left(\frac{1}{N}\boldsymbol{X}\right)\right]$ is an unbiased estimator of $p(\boldsymbol{\pi})$, that is,*

$$E\left[\psi_N\left[p\left(\frac{1}{N}\boldsymbol{X}\right)\right]\right] = p(\boldsymbol{\pi})\ .\square$$

Lemma 2 *If $p(\boldsymbol{x}) = \sum_{i,j} a_{ij} x_i x_j + \sum_i b_i x_i$, then*

$$|p(\boldsymbol{x}) - p(\boldsymbol{y})| \leq (a(\|\boldsymbol{x}\|_1 + \|\boldsymbol{y}\|_1) + b)\|\boldsymbol{x} - \boldsymbol{y}\|_1,$$

where $a = \max\{|a_{ij}|\}$; $b = \max\{|b_i|\}$, and $\|\ \|_1$ denotes the L^1 norm. $\square$

The following result is just a special case of the well-known Azuma–Hoeffding inequality in martingale theory (see for instance, Alon and Spencer [1]).

Lemma 3 *Suppose* $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$ *are independent random variables taking values in any set* S, *and* $L : S^N \to \mathcal{R}$ *is any function that satisfies the condition:*

$$|L(\boldsymbol{u}) - L(\boldsymbol{v})| \leq t$$

whenever $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *differ at just one coordinate. Then,*

$$\mathbb{P}[L(\boldsymbol{X}) - \mathbb{E}[L(\boldsymbol{X})] \geq \lambda] \leq \exp\left(-\frac{\lambda^2}{2Nt^2}\right), \text{ and}$$
$$\mathbb{P}[L(\boldsymbol{X}) - \mathbb{E}[L(\boldsymbol{X})] \leq -\lambda] \leq \exp\left(-\frac{\lambda^2}{2Nt^2}\right) .\square$$

With an eye on our final goal, we immediately describe our method for reconstruction of 4-leaf trees in a setting of reconstructing binary subtrees of given tree T, on which sites developed according to the CF model. Select a quartet $q = \{\alpha, \beta, \gamma, \delta\}$ from $[n]$. For $i, j \in [n]$, $i \neq j$, let $L^{ij} = L^{ij}(\boldsymbol{x})$ be the linear form in indeterminates $\boldsymbol{x} = [x_\chi]$ defined by

$$L^{ij}(\boldsymbol{x}) := \sum_{\chi:\ \chi(i) \neq \chi(j)} x_\chi .$$

Thus, $L^{ij}(\boldsymbol{x}[\boldsymbol{s}])$ is the proportion of sites in the aligned sequences that assign different states to sequences i and j, often called the *dissimilarity score* of sequences i and j. Form the following quadratic polynomials in indeterminates $\boldsymbol{x} = [x_\chi]$:

$$\begin{array}{rcccccc} l^\beta & = & L^{\alpha\beta} & + & L^{\gamma\delta} & - & 2L^{\alpha\beta}L^{\gamma\delta} \\ l^\gamma & = & L^{\alpha\gamma} & + & L^{\beta\delta} & - & 2L^{\alpha\gamma}L^{\beta\delta} \\ l^\delta & = & L^{\alpha\delta} & + & L^{\beta\gamma} & - & 2L^{\alpha\delta}L^{\beta\gamma}. \end{array}$$

Consider the following procedure Φ^1 which inputs a quartet $q \in \binom{[n]}{4}$ and outputs a bph tree $\in B(q)$, i.e. a quartet split of q.

Procedure Φ^1 :

> Given N sites s and a quartet q from $[n]$, set (for $\imath = \beta, \gamma, \delta$),
>
> $$h^\imath(\boldsymbol{s}) := \psi_N[l^\imath(\boldsymbol{x})]|_{\boldsymbol{x}=\boldsymbol{x}[\boldsymbol{s}]},$$
>
> where ψ_N is the linear transformation on $\mathcal{R}_2[\boldsymbol{x}]$ described in (10).
> If $h^\imath(\boldsymbol{s})$ is the (strictly) smallest of $h^\beta(\boldsymbol{s})$, $h^\gamma(\boldsymbol{s})$, $h^\delta(\boldsymbol{s})$ then output the binary tree that groups species $\imath$ with α. In case none of these three numbers is strictly minimal, output `fail`.

We now provide a lower bound on the probability that method Φ^1 returns the correct binary subtree $T^*_{|q}$, i.e. the valid quartet split, for a given sequence length. This bound will be particularly useful when the tree that generated the data has mutation probabilities that are not too small on the internal edge, and not too large on the pendant edges. We may assume w.l.o.g. $q = \{1, 2, 3, 4\}$. Suppose that in $T^*_{|q}$, p_i denotes the mutation probability on the pendant edge incident with leaf i (for $i = 1, 2, 3, 4$), and p_5 denotes the mutation probability on the internal edge.

Theorem 3 *Suppose that, in the underlying four-species tree $T^*_{|q}$ in the CF model,*

$$p_5 \geq \delta, \text{ and}$$
$$p_j \leq (1-\epsilon)/2$$

for $j = 1, 2, 3, 4$, and some $\epsilon, \delta > 0$. If N sites $\boldsymbol{S}$ evolve under the CF model on $T^*_{|q}$, then

$$\mathbb{P}[\Phi^1(\boldsymbol{x}[\boldsymbol{S}]) = T^*_{|q}] \geq 1 - 2\exp(-\beta\delta^2\epsilon^8 N),$$

where $\beta > 0$ is a constant, not dependent on ϵ, δ or N.

Proof. Without loss of generality, suppose that $T^*_{|q}$ is the binary tree that groups together species 1 and 2. For $i = 3, 4$, let

$$R^i = h^i(\boldsymbol{S}) - h^2(\boldsymbol{S}).$$

Then, $\Phi^1(\boldsymbol{x}[\boldsymbol{S}]) = T^*_{|q}$ precisely if R^3 and R^4 are both strictly positive. Thus,

$$\mathbb{P}[\Phi^1(\boldsymbol{x}[\boldsymbol{S}]) = T^*_{|q}] = 1 - \mathbb{P}[\{R^3 \leq 0\} \cup \{R^4 \leq 0\}] \geq 1 - (\mathbb{P}[R^3 \leq 0] + \mathbb{P}[R^4 \leq 0])\ .$$

Simple algebra gives

$$\mathbb{P}[R^i \leq 0] = \mathbb{P}[R^i - \mathbb{E}[R^i] \leq -\mathbb{E}[R^i]] \tag{11}$$

and

$$\mathbb{E}[R^i] = \mathbb{E}[h^i] - \mathbb{E}[h^2] = l^i(\boldsymbol{f}) - l^2(\boldsymbol{f})\ ,$$

by Lemma 1. Now,

$$\begin{aligned} l^2(\boldsymbol{f}) &= \frac{1}{2}\left(1 - \prod_{i=1}^{4}(1-2p_i)\right);\ \text{and} \\ l^3(\boldsymbol{f}) &= l^4(\boldsymbol{f}) = \frac{1}{2}\left(1 - (1-2p_5)^2\prod_{i=1}^{4}(1-2p_i)\right) \end{aligned}$$

since, by (1), $L^{ij}(\boldsymbol{f}) = \frac{1}{2}\left(1 - \prod_{k\in A}(1-2p_k)\right)$ where $\{e_k : k \in A\}$ is the path in $T^*_{|q}$ connecting leaves i and j, and by definition of the l^i. Consequently, for $i = 3, 4$,

$$\mathbb{E}[R^i] = 2p_5(1-p_5)\prod_{j=1}^{4}(1-2p_j) \geq \delta\epsilon^4\ . \tag{12}$$

Combining (11) and (12) we have, for $i = 2, 3$,

$$\mathbb{P}[R^i \leq 0] \leq \mathbb{P}[R^i - \mathbb{E}[R^i] \leq -\delta\epsilon^4]\ .$$

Now, regarding R^i as a function of $S_1, \ldots, S_N$ we see that R^i satisfies the hypothesis of Lemma 3 with $t = \beta'/N$ for some constant $\beta' > 0$. Thus, by Lemma 3, we have

$$\mathbb{P}[R_i \leq 0] \leq \exp(-\beta\delta^2\epsilon^8 N),\ \text{for}\ \beta = 1/(2\beta'^2),$$

as claimed. □

3.3 RECONSTRUCTING N-SPECIES TREES

Suppose a method for constructing a bph tree on four species returns the correct tree with probability $1-\epsilon$ under some model. It is easy to extend such a method to one that constructs a tree on n sequences with high probability, which is both consistent and efficient (i.e. the time required to output a tree grows polynomially with n)—we could simply look at all quartet splits, and if they are consistent with a binary tree, then output this tree, otherwise output the message `fail`. Such a method may require N to grow quickly in order to find the true tree with high probability, and for this reason we wish to avoid using pairs of leaves that are "far apart" in the tree, and thereby likely to mislead tree reconstruction. Thus, we now describe a more refined algorithm that takes account of this.

Consider the following procedure, Φ^*, that, given the dissimilarity score between species, extends procedure Φ^1 for reconstructing a phylogenetic tree from sites for four species, to a procedure that applies to n species.

Procedure Φ^*

Step 1. Define any total order $\leq$ on $\binom{[n]}{4}$ for which:

$$q \leq q' \text{ whenever } \max\{L^{ij}(\boldsymbol{x}[\boldsymbol{s}]) : i,j \in q\} \leq \max\{L^{ij}(\boldsymbol{x}[\boldsymbol{s}]) : i,j \in q'\}.$$

Let Q_i denote the smallest i elements of $\binom{[n]}{4}$ under this ordering. For each $q \in Q_i$, calculate $\Phi^1(q)$. Let $F_i = \{\Phi^1(q) : q \in Q_i\}$.

Step 2. For $i = 1, 2, \ldots$ do:
Compute $cl_2(F_i)$. If $cl_2(F_i) = Q(T)$ for a binary tree T, output T and stop.
If $cl_2(F_i)$ contains a contradictory pair, or if $i = \binom{n}{4}$ output `fail` and stop.
Otherwise return.

We now show that if N sites evolve under the CF model, and the mutation probabilities lie in a certain region, this technique Φ^* requires N to grow sublinearly with n, in order that the correct tree for the n species be recovered from N sites with probability $1-\epsilon$. We first define this "good" region of parameter space for procedure Φ^*. Let $R(n)$ be the interval

$$R(n) = [f(n), g(n)]$$

where $0 < f(n) \leq g(n) < 0.5$, and let $\lambda(n) = (1-2g(n))^{18\log n}$.

Theorem 4 *Suppose N sites evolve under the CF model on $T \in B(n)$, so that for all edges e, $p_e \in R(n)$. Let $N_\epsilon(n)$ denote the smallest number of sites for which $\mathbb{P}[\Phi^*(\boldsymbol{x}[\boldsymbol{S}]) = T] \geq 1-\epsilon$, for fixed $\epsilon \in (0,1)$. Then,*

(1) $N_\epsilon(n) < \dfrac{K \log n}{f^2(n)\lambda^{24}(n)}$
for a constant K.

(2) In particular, $\lim_{n\to\infty} N_\epsilon(n)/n = 0$
if $f(n) = n^{-\alpha}$, *where* $\alpha < 0.5$, *and* $g(n) = \delta(\alpha)$ *is a constant small enough.*

(3) For fixed $k \geq 1$, c, d *constants, if* $f(n) = \dfrac{c}{(\log n)^k}$; $g(n) = \dfrac{d \log\log n}{\log_2 n}$,
then $N_\epsilon(n) \leq (\log n)^{1+2k+864d+o(1)}$.

Proof. Let $\partial_{ij} = \mathbb{E}[\psi_N(L^{ij}(\boldsymbol{x}[\boldsymbol{S}])]$. By Lemma 1, $L^{ij}(\boldsymbol{f}) = L^{ij}(\mathbb{E}[\boldsymbol{x}[\boldsymbol{S}]]) = \partial_{ij}$ is the probability that species i and j are in different states at a site that evolves under the CF model.

For N evolving sites $\mathbf{S}$ and $\tau > 0$, let us define the following three random variables:

$$\begin{aligned} S_\tau &= \{\{i,j\} : L^{ij}(\boldsymbol{x}[\boldsymbol{S}]) < 0.5 - \tau\}, \\ Z &= \{q \in \binom{[n]}{4} : \text{ for all } i,j \in q, \{i,j\} \in S_{2\tau}\}, \text{ and} \\ Z^* &= \{\Phi^1(q) : q \in Z\}. \end{aligned}$$

Also, recall the definition of $D(T)$ and $S(T)$ from Theorem 2: the "short" quartets of T and their quartet splits.

Then Φ^* outputs the correct tree if the following two events A, B occur:

A: $D(T) \subseteq Z$,

B: Φ^1 correctly reconstructs $T^*_{|q}$ for all $q \in Z$,

because, $cl_2(Z^*) \supseteq cl_2(S(T)) = Q(T)$ (by the definition of A and Theorem 2) and $Q(T) \supseteq cl_2(Z^*)$ (by the definition of B) and together these give $cl_2(Z^*) = Q(T)$.

Thus,

$$\mathbb{P}[\Phi^*(\boldsymbol{x}[\boldsymbol{S}]) = T] \geq \mathbb{P}[cl_2(Z^*) = Q(T)] \geq \mathbb{P}[A \cap B].$$

Let C be the event:

$S_{2\tau}$ contains all pairs $\{i,j\}$ with $\partial_{ij} < 0.5 - 3\tau$, and no pair $\{i,j\}$ with $\partial_{ij} \geq 0.5 - \tau$.

We claim that:

$$\mathbb{P}[C] \geq 1 - (n^2 - n)e^{-\tau^2 N/2} \tag{13}$$

and

$$\mathbb{P}[A|C] = 1, \text{ if } \tau \leq \lambda^3(n)/6 \ . \tag{14}$$

To establish (13), first note that $L^{ij}(\boldsymbol{x}[\boldsymbol{S}])$ satisfies the hypothesis of Lemma 3 (with $X_i = S_i$ and $t = 1/N$). Suppose $\partial_{ij} \geq 0.5 - \tau$. Then,

$$\begin{aligned}\mathbb{P}[\{i,j\} \in S_{2\tau}] &= \mathbb{P}[L^{ij}(\boldsymbol{x}[\boldsymbol{S}]) < 0.5 - 2\tau] \leq \mathbb{P}[L^{ij}(\boldsymbol{x}[\boldsymbol{S}]) - \partial_{ij} \leq 0.5 - 2\tau - \partial_{ij}] \leq \\ &\quad \mathbb{P}[L^{ij}(\boldsymbol{x}[\boldsymbol{S}]) - \mathbb{E}[L^{ij}(\boldsymbol{x}[\boldsymbol{S}])] \leq -\tau] \leq e^{-\tau^2 N/2}.\end{aligned}$$

Since there are at most $\binom{n}{2}$ such pairs $\{i,j\}$, the probability that at least one such pair lies in $S_{2\tau}$ is at most $\binom{n}{2} e^{-\tau^2 N/2}$. By a similar argument, the probability that $S_{2\tau}$ fails to contain a pair $\{i,j\}$ with $\partial_{ij} < 0.5 - 3\tau$ is also at most $\binom{n}{2} e^{-\tau^2 N/2}$. These two bounds establish (13).

We start to establish (14). For $q \in D(T)$ and $i,j \in q$, if a path $e_1 e_2 \ldots e_k$ joins leaves i and j, then $k \leq 54 \log n$ by the definition of $D(T)$, and

$$\partial_{ij} = 0.5\left[1 - (1-2p_1)\cdots(1-2p_k)\right] \leq 0.5\left[1 - (1-2g(n))^{54\log n}\right]$$

using $p_e \leq g(n)$ for edges e in T and (1). Thus, $\partial_{ij} < 0.5[1 - \lambda^3(n)]$. Consequently, $\partial_{ij} < 0.5 - 3\tau$ (by assumption that $\tau \leq \lambda^3(n)/6$) and so $\{i,j\} \in S_{2\tau}$ once we condition on the occurence of event C. This holds for all $i,j \in q$, so by definition of Z we have $q \in Z$. This establishes (14).

Set $\tau = \lambda^3(n)/6$. Then for any quartet $q \in D(T)$, the tree $T^*_{|q}$ has mutation probability at least $f(n)$ on its central edge, since $p \geq \min\{p_1, \ldots, p_k\}$ in (1). Furthermore, conditional on C, the mutation probability on any pendant edge is no more than $\max\{\partial_{ij} : i,j \in q\} < 0.5 - \tau = 0.5\left[1 - \frac{\lambda^3(n)}{3}\right]$. Thus, by Theorem 3 for $\epsilon = \tau$ and the Bonferroni inequality,

$$\mathbb{P}[B|C] \geq 1 - 2\binom{n}{4}\exp(-\beta f^2(n)\lambda^{24}(n)N), \tag{15}$$

for a suitable constant $\beta > 0$.

Combining the above, and invoking (14), we have:

$$\mathbb{P}[\Phi^*(\boldsymbol{x}[\boldsymbol{S}]) = T] \geq \mathbb{P}[A \cap B] = \mathbb{P}[A \cap B|C] \times \mathbb{P}[C] = \mathbb{P}[B|C] \times \mathbb{P}[C] \ .$$

¿From (13) and (15) we have:

$$\mathbb{P}[\Phi^*(\boldsymbol{x}[\boldsymbol{S}]) = T] \geq 1 - 2\binom{n}{4}\exp(-\beta f^2(n)\lambda^{24}(n)N) - (n^2 - n)e^{-\lambda^6(n)N/72}$$

and so if we set $N(n) = \frac{C \log n}{f^2(n)\lambda^{24}(n)}$ for a constant C, then we can choose C sufficiently large so that both of the terms involving exponentials decay to zero as n tends to infinity. Now Part (1) holds for a large constant K. Parts (2) and (3) now follow from (1) after some straightforward calculations. □

4 CONCLUSION

(1) A desirable goal would be a tree reconstruction method Φ which satisfies the following three conditions for some suitably small value of ϵ (such as 0.05):

(1) $\Phi(\boldsymbol{x}[\boldsymbol{s}])$ can be constructed by an algorithm whose complexity is polynomial in nN.

(2) The probability that $\Phi(\boldsymbol{x}[\boldsymbol{S}])$ is either the true tree (T_M) or is the message `fail` is at least $1 - \epsilon$ whatever the parameter settings in M.

(3) The probability that $\Phi(\boldsymbol{x}[\boldsymbol{S}])$ is the message `fail` tends to zero as N tends to infinity.

We do not have such a method and do not know if such methods exist at all.

(2) The techniques developed in Subsection 3.2 are likely to be useful in the theory of phylogenetic invariants. A *phylogenetic invariant* for a bph tree T and class ϑ of models M is a polynomial p in variables $\boldsymbol{x} = [x_\chi]$ with $p(\boldsymbol{f}(M)) = 0$, whenever $T_M = T$ and $M \in \vartheta$. For example, the Cavender–Farris model possesses two quadratic polynomial invariants for each binary tree on four leaves, first discovered by Cavender and Felsenstein [7]. Phylogenetic invariants are potentially useful in reconstructing T_M from $\boldsymbol{x}[\boldsymbol{S}]$. The idea is that if p is a phylogenetic invariant for T, then under the assumption that $T = T_M$, the random variable $p(\boldsymbol{x}[\boldsymbol{S}])$ is asymptotically (for N large) normally distributed with mean 0 and a standard deviation that is proportional to $N^{-0.5}$. Thus, if $p(\boldsymbol{x}[\boldsymbol{S}])$ lies too far from 0 for the particular value of N, then one can reject T as a possible candidate for T_M.

However this analysis is asymptotic, and for any particular value of N, and any nonlinear phylogenetic invariant p, the expected value of p differs from 0 (since $\mathbb{E}[p(\boldsymbol{x}[\boldsymbol{S}])] \neq p(\mathbb{E}[\boldsymbol{x}[\boldsymbol{S}]) = p(\boldsymbol{f}) = 0)$. However for any value of N we have the following result whose proof follows directly by combining Lemmas 1, 2, 3.

Proposition 1 Suppose $p(\boldsymbol{x})$ is a quadratic phylogenetic invariant for a model M with underlying tree T. Then, if N sites $\boldsymbol{S}$ evolve under this model, $i(\boldsymbol{S}) := \psi_N[p(\boldsymbol{x}[\boldsymbol{S}])]$ has expected value 0. Furthermore,

$$\mathbb{P}[|i(\boldsymbol{S})| \geq \lambda] \leq 2\exp(-\beta\lambda^2 N),$$

where $\beta > 0$ is a constant dependent only on the coefficients of p. □

(3) We have a result stronger than Theorem 2. Define the *depth* $d(T)$ of a bph tree, which is the maximum distance of any edge from the nearest leaf. Then the valid quartet splits of 4-subtrees in which:
(i) the middle edge is not subdivided, and
(ii) none of the 4 paths representing edges of the 4-subtree is longer than $2d(T)$,
semidyadically determine T. The proof is too lengthy for this paper.

References

[1] N. Alon and J. H. Spencer, *The Probabilistic Method,* John Wiley and Sons, New York, 1992.

[2] H.-J. Bandelt and A. Dress, Reconstructing the shape of a tree from observed dissimulating data, *Adv. App. Math.*, **7**, 309–343 (1986).

[3] D. J. Bryant and M. A. Steel, Extension operations on sets of leaf-labelled trees, *Adv. App. Math.*, **16**, 425–453 (1995).

[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] L. L. Cavalli-Sforza and A. W. F. Edwards, Phylogenetic analysis: models and estimation procedures, *Evolution*, **21**, 550–570 (1967).

[6] J. A. Cavender, Taxonomy with confidence, *Math. Biosci.*, **40**, 271–280 (1978).

[7] J. A. Cavender and J. Felsenstein, Invariants of phylogenies: simple case with discrete states, *J. Classification*, **4**, 57–71 (1987).

[8] J. T. Chang and J. A. Hartigan, Reconstruction of evolutionary trees from pairwise distributions on current species, *Computing science and statistics: Proceedings of the 23rd symposium on the interface*, 254–257 (1991).

[9] G. A. Churcill, A. von Haeseler, and W. C. Navidi, Sample size for a phylogenetic inference, *Mol. Biol. Evol.*, **9(4)**, 735–769 (1992).

[10] H. Colonius and H. H. Schultze, Tree structure for proximity data, *Brit. J. Math. Stat. Psychol.*, **34**, 167–180 (1981).

[11] M. C. H. Dekker, *Reconstruction methods for derivation trees*, Master's Thesis, Vrije Universiteit, Amsterdam, 1986.

[12] P. Erdős and A. Rényi, On a classical problem in probability theory, *Magy. Tud. Akad. Mat. Kutató Int. Közl.*, **6**, 215–220 (1961).

[13] J. S. Farris, A probability model for inferring evolutionary trees, *Syst. Zool.*, **22**, 250–256 (1973).

[14] J. Felsenstein, Cases in which parsimony or compatibility methods will be positively misleading, *Syst. Zool.*, **27**, 401–410 (1978).

[15] J. Felsenstein, Evolutionary trees from DNA sequences: a maximum likelihood approach, *J. Mol. Evol.*, **17**, 368–376 (1981).

[16] N. Goldman, Maximum likelihood inference of phylogenetic trees, with special reference to a poisson process model of DNA substitution and to parsimony analysis, *Syst. Zool.*, **39(4)**, 345–361 (1990).

[17] M. D. Hendy, The relationship between simple evolutionary tree models and observable sequence data, *Syst. Zool.*, **38(4)**, 310–321 (1989).

[18] G. Lecointre, H. Philippe, V. Lé, and H. Le Guyader, How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences, *Mol. Phyl. Evol.*, **2**, 205–224 (1994).

[19] J. Pearl and M. Tarsi, Structuring causal trees, *J. Complexity*, **2**, 60–77 (1986).

[20] H. Philippe and E. Douzery, The pitfalls of molecular phylogeny based on four species, as illustrated by the cetacea/artiodactyla relationships, *J. Mammal. Evol.*, **2(2)**, 133–152 (1994).

[21] N. Saitou, Maximum likelihood methods, *Meth. Enzym.*, **183**, 584–598 (1990).

[22] M. A. Steel, The complexity of reconstructing trees from qualitative characters and subtrees, *J. Classification*, **9**, 91–116 (1992).

[23] M. A. Steel, Recovering a tree from the leaf colourations it generates under a Markov model, *Appl. Math. Lett.*, **7(2)**, 19–24 (1994).

[24] M. A. Steel, L. A. Székely, and M. D. Hendy, Reconstructing trees when sequence sites evolve at variable rates, *Journal of Comput. Biol.*, **1(2)**, 153–163 (1994).

[25] D. L. Swofford and G. J. Olsen, *Molecular Systematics*, Sinaur Associates, Sunderland, 1990.

[26] N. Takezaki and M. Nei, Inconsistency of the maximum parsimony method when the rate of nucleotide substitution is constant, *J. Mol. Evol.*, **39**, 210–218 (1994).

[27] T. Warnow, *Combinatorial algorithms for constructing phylogenetic trees*, PhD thesis, University of California-Berkeley, 1991.

[28] A. Zharkikh and W. H. Li, Inconsistency of the maximum-parsimony method: The case of five taxa with a molecular clock, *Syst. Biol.*, **42**, 113–125 (1993).