TEASING APART TWO TREES

M. A. STEEL AND L. A. SZÉKELY

ABSTRACT. A widely-studied model for generating binary sequences is to 'evolve' them on a tree according to a symmetric Markov process. We show that under this model distinguishing the true (model) tree from a false one is substantially "easier" (in terms of the sequence length needed) than determining the true tree. The key tool is a new and tight Ramsey-type result for binary trees.

1. INTRODUCTION

In this paper we investigate the reconstruction of trees from binary sequences that have been generated on the tree by a simple Markov model. Such processes are widely-studied in molecular genetics, and in other areas of applied probability (including broadcasting and statistical physics). The questions we consider are motivated in part by the concept of NP in computational complexity. In that setting one may have short proofs to the correctness of answers to decision problems that require substantial search. For example, it is hard to find a Hamiltonian cycle in a graph, but if we are given the solution, it is easy to verify. As we pointed out in [5], reconstructing trees from randomly-evolved sequences has two distinct complexity problems. One is the sequence length: what is the minimum sequence length to do phylogeny reconstruction with probability near 1, using *any* algorithm? The other is the more conventional computational complexity of the problem. In this paper we focus on the first problem, the sequence length. In particular we initiate the study of the following natural question: is it possible using *few* observed patterns from the true tree, *to test* an input tree, i.e. reject any false input tree with high probability, and to accept an input true tree with high probability? If yes, can this be done in polynomial time? We make the following progress in this direction: using *few* observed patterns from the true tree, if we are shown the true tree and a false one, we can tell with high probability which one is true, and we make this decision in polynomial time.

In practical applications one might use the maximum likelihood scores of the two trees for this test. However this faces two formidable unsolved problems: (1) computing maximum likelihood scores of trees in polynomial time and (2) showing that maximum likelihood prefers the true tree with high probability from very short

Date: 23 February 2005.
We thank the NZIMA (Maclauren Fellowship) for supporting this research. The second author was supported in part by NSF contracts Nr. 0072187 and 0302307 and NIHGMS contract 1 R01 GM078991-01.

sequences. For (1), see [7, 14]; for (2), see [19]. Accordingly the test we describe here involves an approach that is different to maximum likelihood estimation.

To describe our results more precisely we first provide some terminology concerning trees and random processes on them. In a tree, vertices of degree 1 are called *leaves* and the edges adjacent to them are called *leaf edges*. Other edges are called *internal edges*. A tree is *binary*, if all vertices have degree 1 or 3. Consider a set X of labels. A *phylogenetic X-tree* is a tree, in which leaves are identified with elements of X. We will regard two phylogenetic X-trees as being identical, if there is a graph isomorphism between them, which in addition, if restricted to X, is the identity function of X. If $|X| = n$, then the number of different binary phylogenetic X-trees is $(2n-5)!!$ $(= 1 \times 3 \times 5 \times \cdots \times (2n-5))$ [13]. The *distance* between two vertices in a tree is the number of edges on the unique path connecting them. A *cherry* is a pair of leaves of distance two apart.

For a phylogenetic X-tree $\mathcal{T}$ and subset Y of X, let $\mathcal{T}(Y)$ denote the minimal subtree that connects the leaves in Y, and let $\mathcal{T}|Y$ denote the associated phylogenetic Y-tree obtained from $\mathcal{T}(Y)$ by suppressing vertices of degree 2. Given an edge e of $\mathcal{T}|Y$ let $l(e)$ denote the number of degree two vertices of $\mathcal{T}(Y)$ that were suppressed in forming e. For a phylogenetic tree $\mathcal{T}$ and an interior edge e let $\mathcal{T}/e$ denote the phylogenetic tree obtained by collapsing edge e.

For a phylogenetic X-tree $\mathcal{T}$ and an edge e of it, we speak about the *split* induced by e. The split is an equivalence relation on X induced by "being in one component" after the deletion of e.

In particular, if $\mathcal{T}(Y)$ is binary and Y has four elements, say $Y = \{a, b, c, d\}$ (sometimes called a *quartet*), then $\mathcal{T}|Y$ has a single internal edge, and the removal of this edge from $\mathcal{T}|Y$ partitions the vertices of Y into two -element classes by connectivity. The 3 possible partitions are called *quartet splits* and denoted as follows:

$$ab|cd \quad \text{or} \quad ac|bd \quad \text{or} \quad ad|bc.$$

If $\mathcal{T}|Y = ab|cd$ we say that $\mathcal{T}$ *displays* $ab|cd$, written $ab|_{\mathcal{T}}cd$.

We now describe a model for the evolution of binary sequences on a tree. This model has been described by various authors (and in a range of disciplines including molecular biology, information theory, and physics; for references see [6, 13]). Here we refer to this model as the *CFN model* (short for 'Cavender-Farris-Neyman model'); it has also been referred to in the literature as the 'symmetric binary channel' and the 'symmetric 2-state Poisson model'. The CFN model provides a simple model for the evolution of purine–pyrimidine sequences. The significance of this simple model is, that phenomena shown for the CFN model often extend to more realistic models of sequence evolution.

Suppose we have two states, 0 and 1, and a phylogenetic X-tree $\mathcal{T}$. The CFN model assigns probabilities to the patterns of state of the elements of X as follows. Let us be given $0 < f \leq g < 1/2$ and associate a number p_e $(f \leq p_e \leq g)$ with the edge e called the *substitution probability*. Let ξ_e denote a random indicator variable associated to edge e with $\mathbb{P}[\xi_e = 1] = p_e$, and assume the ξ_e's are independent. Fix any vertex v. For every vertex u, there is a unique path denoted $path(u,v)$ in $\mathcal{T}$. After an experiment on the ξ_e's, define

$$state(u) = state(v) + \sum_{e \in path(u,v)} \xi_e \mod 2. \tag{1}$$

One approach to define *pattern probabilities* is to assign state 0 and 1 to v with probability 1/2, and then compute the probabilities of all possible state patterns of leaves under (1). An essentially equivalent approach is to consider as *pattern* the classes of the equivalence relation "being in the same state" for the elements of X. A pattern of this second kind is the identification of two complementary patterns of the first kind. Let the pattern σ denote a state assignment to every leaf. Let $\mathcal{P}_\sigma$ denote the probability of observing pattern σ under the first approach. Then, if $\bar{\sigma}$ is the probability of the bitwise complementary pattern, then $\mathcal{P}_\sigma = \mathcal{P}_{\bar{\sigma}}$, and if $[\sigma]$ denotes the class of σ after the identification of complementary patterns, then simply $\mathcal{P}_{[\sigma]} = 2\mathcal{P}_\sigma$.

The probability p that the endpoints of a path uw in a CFN tree $\mathcal{T}$ are in different states is nicely related to the substitution probabilities of edges of the uw-path:

$$p = \frac{1}{2}\left(1 - \prod_{e \in path(u,w)} (1 - 2p_e)\right). \tag{2}$$

Formula (2) is well-known, and is easy to prove by induction. Formula (2) also shows that the substitution probability of a path is not less than the smallest transition probability on its edges. It is well-known ([17]) that (1) changing the location of v in $\mathcal{T}$, or (2) substituting a path with internal vertices of degree 2 with a *single edge* in a CFN tree, and assigning to the new edge a transition probability according to (2) *does not change* the probability distribution of patterns.

Usually k independent experiments, called *sites*, are sampled from a binary CFN tree $\mathcal{T}$. The (abstract) *phylogeny reconstruction problem* is the following: from the observed pattern frequencies, tell with a prescribed probability, what was the underlying binary phylogenetic X-tree. We have shown in [5] and [18] that if $|X| = n$ and $n \to \infty$, then $k = \Omega(\log n)$ sites are needed to return the true underlying tree with probability at least $\frac{1}{2} + \epsilon$ with either a deterministic algorithm or with a randomized algorithm whose random bits are independent from the random events on the CFN tree. Sequence length requirements for accurate tree reconstruction is not only of mathematical interest, but also a topical issue in molecular systematics (eg. [2, 12]).

In (most of) this paper we study problems where the bounds f, g are *fixed*, and we let $n \to \infty$. However many of the results generalize to provide results where these quantities are allowed to depend on n, provided the dependence on n does not exceed a certain critical value.

2. Results

The main result of this paper is in Section 4. We show there that for any fixed values of f and g (the minimum and maximum possible substitution probability) with $0 < f \leq g$ and g sufficiently small, *some constant number of sites* generated by the true tree is enough to tell apart the true tree from a false one, if these two trees are given as input. This decision is achieved with near 1 probability, and can be made by a polynomial time algorithm. This fixed number of sites is independent of the number of vertices of the tree (in contrast to the $\Omega(\log n)$ lower bound for the sequence length required for abstract phylogeny reconstruction, mentioned above). The proof of this main result in Section 4 depends on a strange characterization of two binary trees being different. This is a tight Ramsey type result proved in Theorem 3.1, and the proof uses a number of tight estimates for a number of novel extremal problems on binary trees. Many of those may be interesting on their own.

The proof of the main result also requires the existence of a certain kind of phylogenetic tree reconstruction method for 4–leaf CFN trees. Roughly speaking, it must return the true tree with near 1 probability from $O(1)$ length sequences, if p_e is between fixed $0 < f \leq g < 1/2$ for leaf edges and $f \leq p_e < 1/2$ for the backbone edge in the true tree; and in addition, if the true tree is any CFN tree, the method must return it with some fixed probability separated from zero. We conjecture that some natural phylogeny reconstruction methods, like maximum likelihood estimation or Buneman's 4-point condition also satisfy this property, but to formulate a proof looked so inconvenient that we rather invented a new method in Section 6, which has these properties.

3. A Ramsey type result for trees

The following Ramsey type combinatorial result has important statistical implications for distinguishing between two trees using O(1) random characters. The result is Ramsey type in the following sense: no matter how different binary phylogenetic X–trees one tries to make, one still finds some regularity.

Theorem 3.1. *For any two different binary phylogenetic X–trees, $\mathcal{T}_1, \mathcal{T}_2$, at least one of the following must occur:*

(i) *$\mathcal{T}_1$ and $\mathcal{T}_2$ share a cherry, or*
(ii) *there exists a 4-element subset Y of X so that*
   - *$\mathcal{T}_1$ and $\mathcal{T}_2$ induce different quartet splits on Y, and*
   - *each leaf edge e of $\mathcal{T}_1|Y$ has at most four subdividing vertices in $\mathcal{T}_1(Y)$, i.e. we have $l_1(e) \leq 4$.*

Using this theorem, one can readily obtain the following result (Theorem 3.2) which is more conveniently formulated for application in the next Section.

Theorem 3.2. *For any two different binary phylogenetic X–trees, $\mathcal{T}_1, \mathcal{T}_2$ there exists a subset Y of X so that:*

(i) *$\mathcal{T}_1|Y \neq \mathcal{T}_2|Y$.*
(ii) *There are interior edges e_1, e_2 of $\mathcal{T}_1|Y$ and $\mathcal{T}_2|Y$ respectively so that*

$$(\mathcal{T}_1|Y)/e_1 = (\mathcal{T}_2|Y)/e_2.$$

(iii) *among the edges of $\mathcal{T}_1|Y$, the four edges adjacent to e_1 may have each at most four subdividing vertices in $\mathcal{T}_1(Y)$, and other edges different from e_1 are not subdivided at all.*

Such a subset Y can be found in polynomial time from $\mathcal{T}_1, \mathcal{T}_2$.

Note the asymmetry of $\mathcal{T}_1$ and $\mathcal{T}_2$ in the theorems above. We do not see any *obvious* reason why these theorems should work, even for some *large* t replacing $l_1(e) \leq 4$ with $l_1(e) \leq t$. We show first how Theorem 3.1 implies Theorem 3.2.

Proof of Theorem 3.2 (assuming Theorem 3.1).

If $\mathcal{T}_1$ and $\mathcal{T}_2$ have no cherries in common, then the implication is obvious and a quartet guaranteed in Theorem 3.1 can be found by checking all quartets. Otherwise recursively contract common cherries: if $\mathcal{T}_1$ and $\mathcal{T}_2$ share a cherry – say a, b – then replace the cherry both in $\mathcal{T}_1$ and $\mathcal{T}_2$ with a single new leaf (ab), to obtain $\mathcal{F}_1$ and $\mathcal{F}_2$ with $X_F = X \setminus \{a, b\} \cup \{(ab)\}$. Repeat this until the case with $\mathcal{T}_1$ and $\mathcal{T}_2$ have no cherries in common applies. Find an appropriate quartet as above, and then undo the contractions to find the required set $Y \subset X$. Clearly, the steps described can all be implemented in polynomial time. $\square$

Fig. 3 shows that "$t = 4$ subdivisions" is the best possible result in Theorem 3.1 or Theorem 3.2.

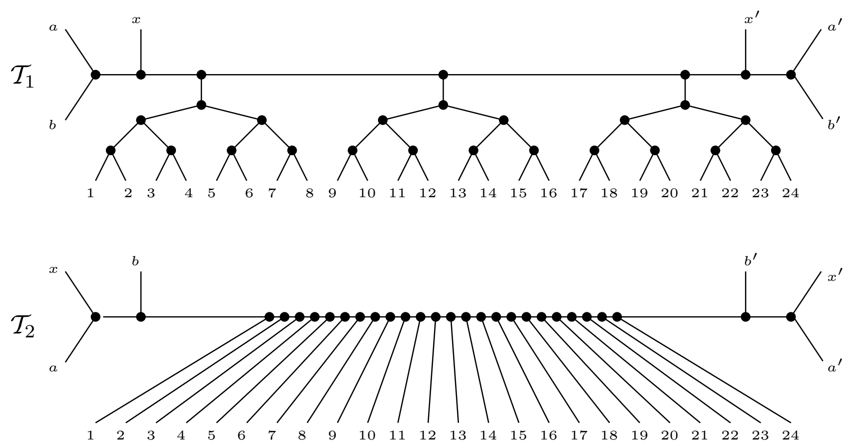

FIGURE 1. Four subdivisions are needed.

Indeed, $Y = \{a, b, x, 8\}$ is, for example, a set satisfying the requirements of the theorem. The path in $\mathcal{T}_1$ between 8 and e_1 of $\mathcal{T}_1|Y$ contains four subdividing vertices. It is easy to see that any other Y satisfying the theorem produces four subdividing vertices in some path as well.

In order to prove Theorem 3.1 we first present a series of lemmas.

Suppose we are given a binary phylogenetic X-tree $\mathcal{T}$ with a positive edge weighting w, and assume a linear order is provided on X. Define the *representative quartet* of an internal edge e of $\mathcal{T}$ to be the quartet (four elements of X) defined as follows. e is adjacent to four edges, e_i, $i = 1, 2, 3, 4$. For every i, consider the shortest (in weighted distance) path containing e_i and some leaf, but not e itself. Select from the leaves realizing the minimum weighted distance from e_i (as above) *the smallest* regarding the linear order given of X. The four leaves obtained in this way make the representative quartet of e. In [5] the following result was established:

Lemma 3.3. *Any binary phylogenetic X-tree $\mathcal{T}$, with any positive edge weighting, and with any linear order given on X, is determined by the set of splits of the representative quartets of $\mathcal{T}$ in the following sense: if all these splits are present in a binary phylogenetic X-tree $\mathcal{T}'$, then $\mathcal{T} = \mathcal{T}'$.*

The proof of the following result (Lemma 3.4) is straightforward, and omitted.

Lemma 3.4. *Assume that $\mathcal{T}$ is a binary tree and leaves u, v form a cherry in $\mathcal{T}$. If vertices c, d are within distance 4 from u (and hence from v), then c and d are within distance 4 from each other.*

Next we introduce a family of equivalence relations on X which will be useful throughout this section. Given a phylogenetic X–tree $\mathcal{T}$ and a positive integer r define a graph on X by joining $i \in X$ and $j \in X$, if their distance in $\mathcal{T}$ is at most r. Consider the transitive closure of this relation, the equivalence relation $\sim_r$. We will generally use $r = 6$ except in Lemma 3.5 and 3.6, where we will also consider $r = 4$.

Lemma 3.5. *Consider a binary X-tree $\mathcal{T}$ with $m = m(\mathcal{T}) = |X| \geq 2$ and the equivalence relation $\sim_4$. Then the equivalence relation has at most $m/2$ classes.*

Proof. The claim is easy to check for $2 \leq m \leq 6$. Assume that $m \geq 7$ and apply induction on m, based on the shape of the ending of a longest path. Fig. 3 shows the cases. $\mathcal{T}_i$ $(i = 1, 2, 3, 4)$ refer to the cases, and $\mathcal{T}_i'$ is the result of truncation as indicated by the curve on Fig. 3.

Let $\#_r(\mathcal{T})$ denote the number of $\sim_r$ classes of the leaf set of the tree T. We have

$$\#_4(\mathcal{T}_1) = \#_4(\mathcal{T}_1') \leq m(\mathcal{T}_1')/2 = m(\mathcal{T}_1)/2 - 1$$

by induction. We claim $\#_4(\mathcal{T}_2) \leq \#_4(\mathcal{T}_2') + 1$. (Indeed, think about the top leaves of $\mathcal{T}_2'$ as u, v in Lemma 3.4. This lemma implies that merging u and v into $\mathcal{T}_2$ and giving up their leafness cannot split any $\sim_4$ class of $\mathcal{T}_2'$ into parts. The top four leaves of $\mathcal{T}_2$ belong to the same $\sim_4$ class of $\mathcal{T}_2$.) Continue with $\#_4(\mathcal{T}_2') + 1 \leq m(\mathcal{T}_2')/2 + 1 = (m(\mathcal{T}_2) - 2)/2 + 1 = m(\mathcal{T}_2)/2$. Similarly,

$$\#_4(\mathcal{T}_3) \leq \#_4(\mathcal{T}_3') + 1 \leq m(\mathcal{T}_3')/2 + 1 = (m(\mathcal{T}_3) - 3)/2 + 1 = m(\mathcal{T}_3)/2 - 1/2,$$

and

$$\#_4(\mathcal{T}_4) \leq \#_4(\mathcal{T}_4') + 1 \leq m(\mathcal{T}_4')/2 + 1 = (m(\mathcal{T}_4) - 4)/2 + 1 = m(\mathcal{T}_4)/2 - 1.$$

□

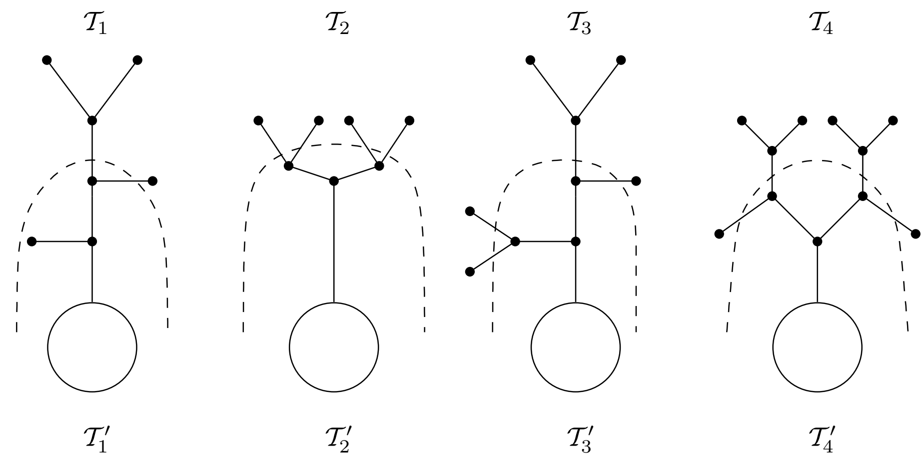

FIGURE 2. Ending of a longest path in a binary tree.

Fig. 3 shows that Lemma 3.5 is basically tight.

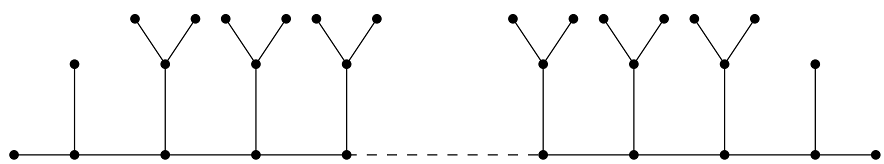

FIGURE 3. $m(\mathcal{T}) = 2k + 4$ and $\#_4(\mathcal{T}) = k$.

Lemma 3.6. *Assume that $\mathcal{T}$ is a binary X-tree on $n \geq 4$ leaves, on which the equivalence relation $\sim_6$ has k classes. Then $\mathcal{T}$ has at least $2k$ cherries.*

Proof. The proof uses induction on n, with base case $n = 4$. We consider two cases: (a) every leaf of $\mathcal{T}$ is in a cherry; (b) $\mathcal{T}$ has a leaf not in a cherry. We will denote by $ch(.)$ the number of cherries, and by $n(.)$ the number of leaves in a tree.
For case (a) replace $\mathcal{T}$ by $\mathcal{T}^*$ contracting every cherry to a single leaf vertex. Then $ch(\mathcal{T}) = n(\mathcal{T}^*) \geq 2\#_4(\mathcal{T}^*)$ by Lemma 3.5, and obviously $\#_4(\mathcal{T}^*) = \#_6(\mathcal{T})$, and so the result holds.
In case (b) assume that x is a leaf in $\mathcal{T}$, but not in a cherry. Then $\mathcal{T}$ must have at least four other leaves. Introduce a new leaf name x' and create two new trees $\mathcal{T}_1$ and $\mathcal{T}_2$ from $\mathcal{T}$ as indicated on Fig. 3.

Both $\mathcal{T}_1$ and $\mathcal{T}_2$ has at least four leaves and the induction hypothesis applies to them. Simple counting shows that

$$ch(\mathcal{T}) = ch(\mathcal{T}_1) + ch(\mathcal{T}_2) - 2. \tag{3}$$

A straightforward case analysis, based on how many of the sets U, V intersects the $\sim_6$-class of x in $\mathcal{T}$, yields that for all cases

$$\#_6(\mathcal{T}) = \#_6(\mathcal{T}_1) + \#_6(\mathcal{T}_2) - 1. \tag{4}$$

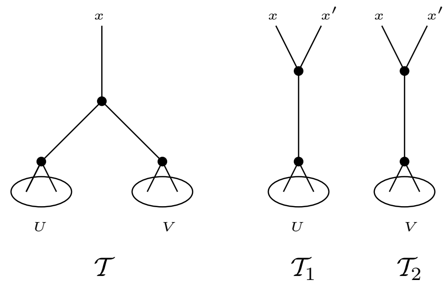

FIGURE 4. Creating two new trees from $\mathcal{T}$ in case (b).

Using (3), the hypothesis for $\mathcal{T}_1$ and $\mathcal{T}_2$, and (4), we obtain

$$ch(\mathcal{T}) = ch(\mathcal{T}_1) + ch(\mathcal{T}_2) - 2 \geq 2\#_6(\mathcal{T}_1) + 2\#_6(\mathcal{T}_2) - 2 = 2\#_6(\mathcal{T}).$$

This completes the proof of Lemma 3.6. □

Proof of Theorem 3.1. Note that Theorem 3.1 holds trivially for $n = 4$. Assume that Theorem 3.1 is not true and let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two phylogenetic X-trees that do not satisfy the conclusions of this theorem. In particular this means that $|X| > 4$ and the following holds:

(5) $\mathcal{T}_1$ and $\mathcal{T}_2$ have no cherries in common,

and

(6) $(Y \subset X,\ |Y| = 4 \text{ and } l_1(e) \leq 4 \text{ for all leaf edge } e \text{ of } \mathcal{T}_1|Y) \implies \mathcal{T}_1|Y = \mathcal{T}_2|Y$

Let $A_1, A_2, ..., A_k$ denote the classes of $\sim_6$ on the tree $\mathcal{T}_1$ (the equivalence relation was defined just prior to Lemma 3.5). We claim that

(7) for $i = 1, 2, ..., k$, $\mathcal{T}_1|A_i$ is isomorphic to $\mathcal{T}_2|A_i$,

and for all $l = 1, 2$ and for all $1 \leq i < j \leq k$, there is a split of $\mathcal{T}_l$ (depending on i, j), such that

(8) A_i and A_j are contained in different sides of the split of $\mathcal{T}_l$.

Once we have established (7) and (8), we have two binary X-trees $\mathcal{T}_1, \mathcal{T}_2$ on n leaves that do not share any cherry (5), and satisfy both (7) and (8). This means that from $\mathcal{T}_l$ $(l = 1, 2)$ one can remove $k - 1$ edges and suppress all vertices of degree 2, to obtain the isomorphic binary forests $\mathcal{F} = \{\mathcal{T}_1|A_i : \ i = 1, 2, ..., k\}$ and $\mathcal{F} = \{\mathcal{T}_2|A_i : \ i = 1, 2, ..., k\}$. Apply now Lemma 3.6 to $\mathcal{T}_1$ to find $2k$ cherries. None of the cherries are divided between the $\sim_6$ equivalence classes, and therefore they are still cherries in the forest $\mathcal{F}$. However, from forest $\mathcal{F}$ we can construct the binary tree $\mathcal{T}_2$ by repeating the following step $k - 1$ times: insert vertices of degree 2 into some edges of the two current components, and join the vertices of degree 2 with a new edge. (If a current component comprises a single vertex, then there is no insertion of a degree 2 vertex, just join the new edge to this single vertex.) Every inserted edge kills at most 2 cherries present, so we can kill at most $2(k - 1)$ cherries of $\mathcal{F}$. Therefore, $\mathcal{T}_1$ and $\mathcal{T}_2$ have a cherry in common, contradicting (5).

Thus all that remains is to establish (7) and (8). This is what comprises the remainder of the proof.

Proof of (7). If $|A_i| \leq 3$, the conclusion holds automatically. Therefore we may assume $|A_i| \geq 4$. We are going to establish that the set of all representative quartet splits of $\mathcal{T}_1|A_i$ are also quartet splits of $\mathcal{T}_2|A_i$. From here, Lemma 3.3 implies $\mathcal{T}_1|A_i = \mathcal{T}_2|A_i$. Consider now an internal edge $f \in E(\mathcal{T}_1|A_i)$. Edge f is the result of contraction of a path π_f in $\mathcal{T}_1(A_i)$. Since A_i was defined as a class of an equivalence relation defined on the vertex set of $\mathcal{T}_1$ by the closure of the relation "exists path of length at most r between 2 leaves", we can think of $\mathcal{T}_1(A_i)$ being covered with paths of length $\leq r$, going from leaf to leaf, and reaching from every leaf every other leaf in a few steps.
To describe the basic quartet split construction, consider four paths in $\mathcal{T}_1(A_i)$, leaving the endvertices u, v of the path π_f, internally disjoint from π_f and each other, connecting to the four closest leaves in $\mathcal{T}_1(A_i)$. Let these leaves be a, b at one end u and c, d at the other end v.
Let $Y = \{a, b, c, d\}$. Observe the following properties of the basic quartet split construction: $d_{\mathcal{T}_1}(a, b) \leq 6$ or ≤ 6-paths passing through u connect a (b) to some leaves. It is easy to see in both cases that $d_{\mathcal{T}_1}(a, u) \leq 5$ and $d_{\mathcal{T}_1}(b, u) \leq 5$. A similar argument yields $d_{\mathcal{T}_1}(c, v) \leq 5$ and $d_{\mathcal{T}_1}(d, v) \leq 5$. Since we just showed that in $\mathcal{T}_1|Y$ for any leaf edge e we have $l_1(e) \leq 4$, (6) implies that $\mathcal{T}_1|Y = \mathcal{T}_2|Y$, with other words $ab|cd$ is displayed by $\mathcal{T}_2$. Next, we show that the quartet splits of the basic quartet split construction in $\mathcal{T}_1(A_i)$ (that a fortiori generates the same quartet splits in $\mathcal{T}_1|A_i$) provide *all* representative quartet splits of $\mathcal{T}_1|A_i$ under some weighting of the edges of $\mathcal{T}_1|A_i$, and therefore we are in a position to apply Lemma 3.3. Namely, use for the weight of an edge of $\mathcal{T}_1|A_i$ the number of edges in the path representing this edge in $\mathcal{T}_1(A_i)$. Indeed, take now any path from $e \in E(\mathcal{T}_1|A_i)$ in a direction to a closest leaf in $\mathcal{T}_1|A_i$. If the weight of this path is w, then there is a path from π_e in $\mathcal{T}_1(A_i)$ to a leaf of distance w, and this must be a closest path in $\mathcal{T}_1(A_i)$. (Any linear ordering of X suffices.) This establishes (7).

Proof of (8). If $|A_i| = 1$ or $|A_j| = 1$, then there is nothing to prove. We may assume $|A_i| \geq 2$ and $|A_j| \geq 2$. Take $a, b \in A_i$ and $c, d \in A_j$ distinct leaves. First we are going to show that $\mathcal{T}_1$ displays the quartet split $ab|cd$, that is:

$$ab|_{\mathcal{T}_1}cd \tag{9}$$

Formula (9) is clearly equivalent to the fact that the ab and cd paths in $\mathcal{T}_1$ are disjoint. Assume for the contrary, that ab and cd paths in $\mathcal{T}_1$ intersect, say edge f is in the intersection.
Recall that the ab path can be covered in $\mathcal{T}_1$ with a sequence of paths of length ≤ 6, each connecting vertices from A_i, and similarly, the cd path can be covered in $\mathcal{T}_1$ with a sequence of paths of length ≤ 6, each connecting vertices from A_j. Some of these ≤ 6-paths contain f, and it follows that there are $a', b' \in A_i$ with $d_{\mathcal{T}_1}(a', b') \leq 6$, and $c', d' \in A_j$ with $d_{\mathcal{T}_1}(c', d') \leq 6$, such that the $a'b'$ and $c'd'$ paths in $\mathcal{T}_1$ intersect. Therefore $a'c'|_{\mathcal{T}_1}b'd'$ or $a'd'|_{\mathcal{T}_1}b'c'$. In the first case, $d_{\mathcal{T}_1}(a', c') + d_{\mathcal{T}_1}(b', d') < 12$, which implies that a', c', b', d' are all in one equivalence class. The second case provides a similar contradiction. Now formula (9) implies that in $\mathcal{T}_1$ any two paths connecting A_i to A_j intersect. Recall the Helly property of trees: if a collection of subtrees of a tree pairwise intersect, then they all intersect (see

[8], Ex. 6.16, p. 41). Therefore all these paths intersect in an edge or a vertex of $\mathcal{T}_1$. A simple case analysis shows that a trivalent vertex cannot be such a common intersection, if it is not part of an intersection edge. This establishes (8) for $\mathcal{T}_1$. For $\mathcal{T}_2$, we still have $|A_i| \geq 2$ and $|A_j| \geq 2$. Let $a, b \in A_i$ with $d_{\mathcal{T}_1}(a, b) \leq 6$ and $c, d \in A_j$ with $d_{\mathcal{T}_1}(c, d) \leq 6$. These with (9) means that for $Y = \{a, b, c, d\}$ for any leaf edge e in $\mathcal{T}_1|Y$ we have $l_1(e) \leq 4$; therefore (6) gives us that

$$ab|_{\mathcal{T}_2} cd \tag{10}$$

We now use the following quartet inference rule: If any binary X-tree displays the quartet splits $ax|cd$ and $bx|cd$ then it necessarily displays the quartet split $ab|cd$ ([3, 4]). Applying this inference rule we see that (10) holds for $a, b \in A_i$ and $c, d \in A_j$ *with no distance condition* as well.

This completes the proof of (8) and thereby of Theorem 3.1. □

4. Telling the true tree apart from a false one

The main result of this paper is the following theorem:

Theorem 4.1. *Suppose we have a true CFN binary phylogenetic X-tree $\mathcal{F}_1$ with transition mechanism $0 < f \leq p_e < \gamma < 1/8$ and another binary phylogenetic X-tree $\mathcal{F}_2$. For every $\epsilon' > 0$ there is a test, which from the input of $K = K(\epsilon', f, \gamma)$ sites evolved on the true tree, and the input of the 2-element set $\{\mathcal{T}_1, \mathcal{T}_2\}$ (where $\{\mathcal{T}_1, \mathcal{T}_2\} = \{\mathcal{F}_1, \mathcal{F}_2\}$), tells which input tree is the true tree with probability $\geq 1 - \epsilon'$. Moreover, this test can be realized by an algorithm that is polynomial time in $|X|$.*

The crucial point is that $K(\epsilon', f, \gamma)$ is *not dependent* on $|X|$.

In order to describe our test we first discuss *ancestral state reconstruction*. Assume that we have binary phylogenetic X-tree. *Subdivide* an edge with a vertex r, and use this r and a character associated with it to build a CFN model with leaf states assigned, as in (1). We call this a *rooted* CFN tree. The *ancestral state reconstruction problem* is as follows: from the states assigned to the leaves, try to identify the state of r (the location of r is known). The *Fitch–Hartigan algorithm* (see [13], p. 90) recursively assigns *state sets* to vertices of the rooted CFN tree, starting with leaves, going towards the root r. If the state of a leaf is x ($x = 0$ or 1), then let the state set of the leaf be $\{x\}$. If both sons of a vertex already have state sets M and N assigned, then assign to this vertex state set $M \cap N$, if $M \cap N \neq \emptyset$, otherwise $M \cup N$. (The original Fitch–Hartigan algorithm was designed to evaluate the *parsimony score* of a given tree, and upon parsing the tree once more from root towards the leaves it also constructs a most parsimonious reconstruction of states on the tree. We use the algorithm here for a different purpose). We will use the following simple algorithm (called *randomized FH*) for ancestral state reconstruction: if the root has a singleton state set, we reconstruct for ancestral state the element of this singleton set; and if the root has a doubleton state set, we select the ancestral state by tossing a fair coin.

The proof of the following result is given in Section 5.

Lemma 4.2. *Consider any rooted binary phylogenetic tree, $\mathcal{T}$ and a binary character evolved under the CFN model with $p(e) < 1/2$ for each edge e of $\mathcal{T}$. Then the probability that the randomized Fitch–Hartigan algorithm correctly reconstructs the root state is*

(i) *strictly greater than $\frac{1}{2}$;*
(ii) *at least $\frac{1}{2} + \Delta_g$, where*

$$\Delta_g = \frac{\sqrt{(1-4g)(1-8g)}}{2(1-2g)^2}, \tag{11}$$

provided that $p_e \leq g < 1/8$ for each edge e of $\mathcal{T}$.

The test that works for Theorem 4.1 is the following. Use Theorem 3.2 to find a set $Y \subset X$ for the ordered pair $\mathcal{T}_1, \mathcal{T}_2$. Consider the tree $\mathcal{T}_1|Y$. Remove the edge e_1 and 4 edges adjacent to it in $\mathcal{T}_1|Y$. We are left with 4 rooted phylogenetic trees, $\mathcal{C}_i$ with root r_i. (Note that $\mathcal{C}_i$ is present in $\mathcal{T}_1$ itself as one side of an edge.) Let the leaf set of $\mathcal{C}_i$ be X_i. Observe that removing the edge e_2 and 4 edges adjacent to it in $\mathcal{T}_2|Y$ results in the very same $\mathcal{C}_i$, X_i, and r_i. Assume without loss of generality that r_1r_2 and r_3r_4 are separated by e_1 in $\mathcal{T}_1|Y$, but r_1r_3 and r_2r_4 are separated by e_2 in $\mathcal{T}_2|Y$.

Consider now an input site developed on the true tree $\mathcal{F}_1$. For $i = 1, 2, 3, 4$, use the randomized Fitch–Hartigan algorithm to reconstruct the ancestral state of r_i using the character states in X_i and the rooted tree $\mathcal{C}_i$. For a K (to be specified later), repeat the ancestral state reconstruction above for K independent sites. We have 4 sequences of length K associated with r_i, $i = 1, 2, 3, 4$. Feed these sequences as an input into a tree reconstruction which meets the specification of the following Lemma.

Lemma 4.3. *Let δ be any positive number that is less than $\frac{1}{3}$. Then there exists a tree reconstruction method $M(= M_\delta)$ for 4-leaf binary phylogenetic trees equipped with the CFN model, which has the following two properties:*

(i) *Suppose that each pendant edge of the true 4-leaf tree has $0 < f \leq p_e \leq g < 1/2$, and we have $0 < f \leq p_e < 1/2$ on the central edge. Then for any $\epsilon > 0$, there exists a constant $k = k(\epsilon, f, g)$ number so that the probability the method M returns the true tree from at least k i.i.d. sites, is at least $1 - \epsilon$.*
(ii) *M correctly reconstructs each true tree $\mathcal{T}$ with probability at least δ under any CFN transition mechanism on $\mathcal{T}$, from any number of sites.*

(Section 6 contains the proof of Lemma 4.3. We conjecture that many tree reconstruction methods, including the Buneman four-point condition and maximum likelihood estimation, actually satisfy these conditions.) We select $\delta = 1/4$, $\epsilon = 0.01$, f and γ from the conditions of Theorem 4.1, g from

$$g = \frac{1}{2}\left(1 - (1-2\gamma)^5\left(1 - 2\left(\frac{1}{2} - \Delta_\gamma\right)\right)\right), \tag{12}$$

and we set M for these numbers. Apply M to $K/k(\epsilon, f, g)$ disjoint $k(\epsilon, f, g)$-tuples of the K sites defined by the 4 sequences associated with r_i ($i = 1, 2, 3, 4$). The output of each use M is a 4-leaf tree, that we identify with one of $r_1r_2|r_3r_4$, $r_1r_3|r_2r_4$, $r_1r_4|r_2r_3$. If we get $r_1r_2|r_3r_4$ in at least 7/8 of the $K/k(\epsilon, f, g)$ outputs of M, we output $\mathcal{T}_1$ as our guess for the true tree $\mathcal{F}_1$. Otherwise we output $\mathcal{T}_2$.

This is the test, and the remainder of this section is devoted to establishing the proof of its correctness. The proof considers two cases: either $\mathcal{T}_1$ is the true tree or $\mathcal{T}_2$ is the true tree.

We select $K = \max(K_1, K_2)$, where $K_1 = K_1(\epsilon', f, \gamma)$ sites are sufficient when $\mathcal{T}_1$ is the true tree and $K_2 = K_2(\epsilon', f, \gamma)$ sites are sufficient when $\mathcal{T}_2$ is the true tree (the K_i will be specified later).

Assume first that $\mathcal{T}_1$ is the true tree, i.e. $\mathcal{T}_1 = \mathcal{F}_1$. Then $\mathcal{C}_i$ has an inherited rooted CFN structure from the true tree, keeping transition probabilities from edge to edge, and γ is an upper bound for these transition probabilities. By Lemma 4.2(ii), the randomized Fitch–Hartigan algorithm applied to $\mathcal{C}_i$ reconstructs the true state of r_i in $\mathcal{F}_1|Y$ with probability at least $\frac{1}{2} + \Delta_\gamma$. Let q_i denote the probability that for any given site the randomized Fitch–Hartigan algorithm incorrectly reconstructs the state of r_i. Let us now equip the 4-leaf binary tree $\mathcal{R}$ identified by $r_1r_2|r_3r_4$ with a transition mechanism as follows. The transition probability of the backbone edge is the transition probability on the path in $\mathcal{T}_1$ corresponding to e_1 in $\mathcal{T}_1|Y$; and if w_i is the transition probability on the path in $\mathcal{T}_1$ corresponding to the edge connecting r_i to e_1 in $\mathcal{T}_1|Y$, let the transition probability p_i of the r_i leaf edge in $\mathcal{R}$ be recomputed by

$$p_i = \frac{1}{2}\Bigg(1 - (1 - 2w_i)(1 - 2q_i)\Bigg) \,. \tag{13}$$

Lemma 4.4. (i) *$\mathcal{R}$ is a CFN model tree with the transition mechanism above;*
(ii) *all edge transition probabilities are at least f (f is from the conditions of Theorem 4.1); and all leaf edge transition probabilities are at most g, where g is as in (12);*
(iii) *the distribution of the leaf coloration pattern of $\mathcal{R}$ is exactly the same as that of the result of the ancestral state reconstruction for r_i ($i = 1, 2, 3, 4$).*

Proof. For part (i) note that independence of transitions on different edges follows from the fact random events influencing transitions on edges of $\mathcal{R}$ come from disjoint edge sets of $\mathcal{T}_1$. It is clear that transition probabilities are in $(0, .5)$. For part (ii), the transition probabilities stay above f, since every edge that may come into a path had transition probability at least f. Clearly $g < 1/2$ from (12). A leaf edge in $\mathcal{R}$ has transition mechanism which a combination of those of at most five edges of $\mathcal{T}_1$ (Theorem 3.2), and of ancestral site reconstruction, which errs with probability $q_i \leq 1/2 - \Delta_\gamma$. From here, the upper bound for the combined transition probability (12) easily follows from (2) and (13). Part (iii) is obvious from the construction. □

When we apply M to a $k(\epsilon, f, g)$-tuple of sites in the test, it is no different from applying M to a $k(\epsilon, f, g)$-tuple of sites of $\mathcal{R}$, according to Lemma 4.4(iii).

Lemma 4.4(i),(ii) certifies that Lemma 4.3 can be used. Therefore M correctly reconstructs $\mathcal{R}$ with probability at least $1-\epsilon = 0.99$. We apply M to $k(\epsilon, f, g)$-tuples of sites $K_1/k(\epsilon, f, g)$ times. The number of cases when $\mathcal{R}$ is correctly returned follows binomial distribution. The probability of getting $\mathcal{R}$ in less than 7/8 of the experiments goes to zero, as $K_1 \to \infty$. For this case of the proof, we can take a $K_1 = K_1(\epsilon', f, \gamma)$ which puts this probability below ϵ'.

Consider now the second case that $\mathcal{T}_2$ is the true tree, i.e. $\mathcal{T}_2 = \mathcal{F}_1$. Then $\mathcal{C}_i$ has an inherited rooted CFN structure from the true tree by suppressing non-root vertices of degree 2 and computing compound transition probabilities by (2) for the new edges. By Lemma 4.2(i), the randomized Fitch–Hartigan algorithm applied to $\mathcal{C}_i$ reconstructs the true state of r_i in $\mathcal{F}_1|Y$ with a probability $> 1/2$. Let q_i denote the probability that in a site the randomized Fitch–Hartigan algorithm incorrectly guessed the state of r_i. Let us know equip the 4-leaf binary tree $\mathcal{R}'$ identified by $r_1r_3|r_2r_4$ with a CFN transition mechanism as follows. The transition probability of the backbone edge is the inherited transition probability of e_2 in $\mathcal{T}_2|Y$; and if w_i is the inherited transition probability of the edge connecting r_i to e_2 in $\mathcal{T}_2|Y$, let the transition probability p_i of the r_i leaf edge be given by (13) again. Now the trick comes again: it is easy to see—we skip the details—that for $\mathcal{R}'$, Lemma 4.4(i), (iii) still hold. When we apply M to a $k(\epsilon, f, g)$-tuple of sites in the test, it is no different from applying M to a $k(\epsilon, f, g)$-tuple of sites of $\mathcal{R}'$, according to Lemma 4.4(iii). Lemma 4.4(i) certifies that Lemma 4.3(ii) can be used. Therefore M correctly reconstructs $\mathcal{R}'$ with probability at least 1/4. We apply M to $k(\epsilon, f, g)$-tuples of sites $K_2/k(\epsilon, f, g)$ times. The number of cases when $\mathcal{R}$ is correctly returned follows binomial distribution. The probability of getting $\mathcal{R}'$ in more than 7/8 of the experiments goes to zero, as $K_2 \to \infty$. For this case of the proof, we can take $K_2 = K_2(\epsilon', f, \gamma)$ which puts this probability below ϵ'.

Remark An interesting question is whether Theorem 4.1 can be improved by allowing a larger bound on γ. It seems possible that one may be able, with more work, to replace the value $\frac{1}{8}$ by a value closer to (or equal to) $\frac{1}{2}(1-\frac{1}{\sqrt{2}})$ which would be the largest possible value for which the Theorem 4.1 can hold (by [6]).

5. Proof of Lemma 4.2

Let S (respectively D) be the probability that the randomized Fitch–Hartigan algorithm applied to a character randomly generated on $\mathcal{T}$ finds a unique state for the root of $\mathcal{T}$ without tossing a coin, and this state is the true state (resp. not the true state). Let E be the probability that the randomized Fitch–Hartigan algorithm applied to a character randomly generated on T finds a doubleton state set at the root, and makes a decision with tossing a coin. Note that the probability that we can reconstruct the true root state using the randomized Fitch–Hartigan algorithm is

$$S + \frac{1}{2}E = \frac{1}{2} + \frac{1}{2}(S-D), \tag{14}$$

(since $S + D + E = 1$). We first show(assuming that the $p(e)$'s are all less that $\frac{1}{2}$) that $S - D > 0$ which implies that $S + \frac{1}{2}E > \frac{1}{2}$ as claimed.

To establish $S - D > 0$ we will use induction on the depth h of $\mathcal{T}$ (the length of the longest path from the root to any leaf). For $h = 1$ the result is easily verified. Suppose it holds for all trees of depth h, and let $\mathcal{T}$ be a tree of depth $h+1$. Consider the two rooted subtrees of T that are incident with the root of $\mathcal{T}$ – call them $\mathcal{T}_1$ and $\mathcal{T}_2$ (one may be a leaf, but this causes no problem). Let the two edges that connect their roots to the root of $\mathcal{T}$ be (respectively) e_1 and e_2, and for $i = 1, 2$ let $p_i := p(e_i), q_i := 1 - p(e_i)$, and by assumption we have $q_i - p_i > 0$.

Now consider computing S, D, E for $\mathcal{T}_i$ (ignoring the rest of $\mathcal{T}$) – call them S_i, D_i, E_i – for $i = 1, 2$. Thus, for example, S_1 is the probability that the randomized Fitch–Hartigan algorithm applied to a character randomly generated on $\mathcal{T}_1$ finds a unique state for the root of $\mathcal{T}_1$ without tossing a coin, and this state is the true state. The following fundamental recursions are from [15]:

$$S = (q_1S_1 + p_1D_1)(q_2S_2 + p_2D_2) + E_1(q_2S_2 + p_2D_2) + E_2(q_1S_1 + p_1D_1),$$

and

$$D = (p_1S_1 + q_1D_1)(p_2S_2 + q_2D_2) + E_1(p_2S_2 + q_2D_2) + E_2(p_1S_1 + q_1D_1).$$

If follows by easy algebra that

$$S-D = (q_1q_2-p_1p_2)(S_1S_2-D_1D_2)+E_1(q_2-p_2)(S_2-D_2)+E_2(q_1-p_1)(S_1-D_1). \tag{15}$$

But each of these three terms is strictly positive, since $q_i > p_i$ and since (by induction) $S_i > D_i$ for $i = 1, 2$. This completes the induction step that $S - D > 0$ and thereby the proof of part (i) of Lemma 4.2.

Part (ii) of Lemma 4.2 was proved in the Honours Thesis of Kahn Mason [9], under the guidance of the first author. Here we provide a shorter, more direct proof.

By a *complete binary tree (of depth h)* we mean a rooted binary tree, having 2^h leaves, each at distance h from the root. Suppose we are given any rooted binary phylogenetic tree $\mathcal{T}$, with substitution probabilities, and for which the maximum distance of any leaf from the root is h. We can convert T into a complete binary tree of depth $h' \geq h$ (with associated substitution probabilities) by the following procedure. If leaf x of $\mathcal{T}$ has distance h_x from the root, then identify x with the root of a complete binary tree of depth $h' - h_x$, and assign a substitution probability 0 to all the new edges in this subtree. If we do this for each leaf, then the resulting tree $\mathcal{T}'$ is a complete binary tree of depth h' and for which $0 \leq p(e) \leq g < 1/8$. Furthermore, the randomized Fitch–Hartigan algorithm correctly reconstructs the true ancestral state of $\mathcal{T}'$ (with substitution probabilities as specified) with exactly the same probability as it does for $\mathcal{T}$.

Now (15) implies that $S - D$ is a strictly decreasing function of p_i and a strictly increasing function of $(S_i - D_i)$ (for $i = 1, 2$). Thus the $S - D$ value for $\mathcal{T}$ is at least the $S - D$ value for $\mathcal{T}'$ with the substitution probabilities as described, and therefore at least the $S - D$ value of a complete binary tree of depth h' having substitution probability equal to g on every edge. Furthermore this holds for all h' greater or equal to the maximal distance of any leaf of $\mathcal{T}$ to the root of $\mathcal{T}$. Now Theorem 3 of [16] whose proof appears in [15] (related but more general results appear in [10]) states that for the (rooted) complete binary tree of depth h and substitution probability $g < \frac{1}{8}$ on every edge, the limiting value (as $h \to \infty$) of $\frac{1}{2}(S-D)$ is Δ_g. Thus Δ_g is a lower bound to $\frac{1}{2}(S-D)$ for $\mathcal{T}$ with its original substitution probabilities, as claimed. This completes the proof of Lemma 4.2.

6. Proof of Lemma 4.3

We first establish two preliminary results. The proof of the first will require the Azuma–Hoeffding inequality (see [1]) which states the following:

Lemma 6.1. *Suppose* $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ *are independent random variables taking values in any set* S, *and* $L : S^k \to \mathbb{R}$ *is any function that satisfies the condition:* $|L(\mathbf{u}) - L(\mathbf{v})| \leq t$ *whenever* $\mathbf{u}$ *and* $\mathbf{v}$ *differ at just one coordinate. Then,*
(16)

$$\mathbb{P}\Bigg[L(\mathbf{X}) - \mathbb{E}[L(\mathbf{X})] \geq \lambda\Bigg] \quad \textit{and} \quad \mathbb{P}\Bigg[L(\mathbf{X}) - \mathbb{E}[L(\mathbf{X})] \leq -\lambda\Bigg] \leq \exp\left(-\frac{\lambda^2}{2t^2k}\right). \qquad \square$$

Lemma 6.2. *Suppose that* $\hat{s}$ *is the frequency counts of a multinomial distribution with* k *trials and with expectation vector* $\mathbb{E}[\hat{s}] = s$. *Then, for* $b > 0$,

$$\mathbb{P}\Bigg[|s - \hat{s}|_2 > (1+b)/\sqrt{k}\Bigg] \leq \exp(-b^2/4).$$

where $|.|_2$ *denotes Euclidean distance.*

Proof. Consider the random variable $Y (= Y(\hat{s})) := |s - \hat{s}|_2$ as a function of the k (independent) site patterns. Suppose one of these site patterns is changed – in which case $\hat{s}$ changes in two co-ordinates by $\frac{1}{k}$ – all other co-ordinates are unchanged. Let $\hat{s}'$ denote this perturbation of $\hat{s}$, and $Y' = Y(\hat{s}')$. Then, by the triangle inequality, $|Y - Y'| \leq |\hat{s} - \hat{s}'|_2 = \frac{\sqrt{2}}{k}$. Consequently, by the Azuma–Hoeffding inequality (Lemma 6.1) we have

$$\mathbb{P}\Bigg[Y - \mathbb{E}[Y] > \lambda\Bigg] \leq \exp(-\lambda^2 k/4).$$

Now, if N denotes the number of categories of the multinomial distribution, then $\mathbb{E}[Y] \leq \sqrt{\mathbb{E}[Y^2]}$ and

$$\mathbb{E}[Y^2] = \mathbb{E}\left[\sum_{i=1}^{N} (\hat{s_i} - s_i)^2\right] = \sum_{i=1}^{N} \mathbb{E}[(\hat{s_i} - s_i)^2] = \sum_{i=1}^{N} \frac{1}{k} s_i(1 - s_i)$$

(here we use the fact that for the frequency counts of a multinomial distribution $\mathbb{E}[(\hat{s_i} - s_i)^2] = Var[\hat{s_i}] = \frac{1}{k} s_i(1-s_i)$.) In particular, $\mathbb{E}[Y^2] \leq \frac{1}{k}$ and so $\mathbb{E}[Y] \leq \frac{1}{\sqrt{k}}$. Consequently,

$$\mathbb{P}\left[Y > \lambda + \frac{1}{\sqrt{k}}\right] \leq \exp(-\lambda^2 k/4).$$

The result now follows by taking $\lambda = b/\sqrt{k}$. $\square$

Lemma 6.3. *Consider the CFN model on two different four taxon trees $\mathcal{T}$ and $\mathcal{T}'$, and suppose that on $\mathcal{T}$ we have the following restriction on the substitution probabilities: for the four leaf edges*

$$0 \leq p_e \leq g < \frac{1}{2}$$

while for the central edge

$$0 < f \leq p_e \leq \frac{1}{2}.$$

Suppose that on $\mathcal{T}'$ the substitution probabilities are completely unrestricted—that is for this tree the only requirement for each edge e is that $0 \leq p_e \leq \frac{1}{2}$. Let s and s' be the vector of probabilities of site patterns produced by $\mathcal{T}$ and $\mathcal{T}'$ (respectively) with substitution probabilities satisfying the constraints described. Then

$$|s - s'|_2 \geq cf(1-2g)^4$$

for an absolute constant $c > 0$ that is independent of f and g.

Proof. Without loss of generality we may assume $X = \{1,2,3,4\}$ and that $\mathcal{T}$ is the tree $12|34$ and $\mathcal{T}'$ is the tree $13|24$. We may regard s and s' as sitting in the 7–dimensional simplex $\Delta_7 \subset \mathbb{R}^8$. Let F be any real-valued differentiable function on this simplex, for which the first derivative of F is bounded above on Δ_7 in absolute value—say by M. Then, by elementary calculus,

$$|F(s) - F(s')| \leq M \sum_{i=1}^{8} |s_i - s'_i| = M|s - s|_1,$$

where $|.|_1$ denotes the L_1–norm. It follows that $|s - s'|_2 \geq \frac{1}{M\sqrt{8}}|F(s) - F(s')|$.

We are going to chose a quadratic function F, and for such a function, a finite positive value of M certainly exists. Thus it suffices to show that

$$|F(s) - F(s')| \geq c' f(1-2g)^4 \tag{17}$$

for a constant c'. Now, let $p(13;24)$ be the probability leaves 1 and 3 are in different states, and leaves 2 and 4 are in different states. Thus $p(13;24)$ is a sum of certain s values (and also a sum of certain s' values). Similarly, let $p(13)$ be the probability that leaves 1 and 3 are in different states, and let $p(24)$ be the probability that leaves 2 and 4 are in different states. These are also linear functions of s (and of s'). Our quadratic function is $F = p(13;24) - p(13)p(24)$. It is well known that $F(s') = 0$ (i.e. F is a quadratic phylogenetic invariant for the tree $13|24$ under the CFN model which reflects the property of that model that changes across two edge-disjoint paths in the tree are statistically independent). On the other hand, algebraic manipulation shows that

$$F(s) = p_0(1-p_0)\prod_{i=1}^{4}(1-2p_i)$$

where p_0 is the substitution probability on the central edge of $\mathcal{T}$, and p_i is the transition probability for the leaf edge of $\mathcal{T}$ incident with i. Consequently, by the restrictions imposed on the p_e values, we have $F(s) \geq f(1-f)(1-2g)^4$, and so we can take $c' = \frac{1}{2}$ in (17). This completes the proof. $\square$

We turn to the proof of Lemma 4.3. Denote the three trees on four taxa by $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$. The method $M = M_\delta$ is as follows. Firstly let b_k be an unbounded increasing sequence, with $\delta = \frac{1}{3}(1-\exp(-b_1/4))$ and $b_k/\sqrt{k} \rightarrow 0$ as $k \rightarrow \infty$.

Given $\hat{s}$ we say one of these three trees – say $\mathcal{T}_i$ - is 'good' (for $\hat{s}$) if we can find substitution probabilities (the p_e values) for $\mathcal{T}_i$ so that $s(\mathcal{T}_i, \{p_e\})$ lies within Euclidean distance at most $(1+b_k)/\sqrt{k}$ from $\hat{s}$ (here $s(\mathcal{T}_i, \{p_e\})$ is the probability vector of patterns generated by $\mathcal{T}_i$ with parameters $\{p_e\}$).

So from $\hat{s}$ we construct the good trees (there may be none, one, two or three). If there are none, select one of the three trees uniformly at random. Otherwise if the set of good trees is nonempty then select one of them uniformly at random.

We claim that M satisfies conditions (i) and (ii) of Lemma 4.3. We suppose without loss of generality that $\mathcal{T}_1$ is the true tree.

To establish condition (i) we need to specify the function $k(\epsilon, f, g)$. Given ϵ, f, g let $k(\epsilon, f, g)$ be the smallest value of k for which the following two inequalities hold:

$$\exp(-b_k^2/4) \leq \epsilon, \tag{18}$$

and

$$\frac{1+b_k}{\sqrt{k}} < \frac{1}{2}cf(1-2g)^4, \tag{19}$$

where c is the constant in Lemma 6.3. Suppose now that $k \geq k(\epsilon, f, g)$ and consider the event E that $\mathcal{T}_1$ is a good tree. Then Lemma 6.2 and condition (18) implies that E has probability at least $1-\epsilon$. Furthermore, if s' is the vector of pattern probabilities generated by one of the other trees, then by the triangle inequality we have:

$$|s' - \hat{s}|_2 \geq |s' - s|_2 - |s - \hat{s}|_2 \geq cf(1-2g)^4 - |s - \hat{s}|_2, \tag{20}$$

where the second inequality is due to Lemma 6.3. Now, conditional on event E, by Lemma 6.2, we have $|s-\hat{s}|_2 \leq \frac{1+b_k}{\sqrt{k}}$ and so, by inequality (19) we have $|s-\hat{s}|_2 < \frac{1}{2}cf(1-2g)^4$. Applying this to (20) and again invoking inequality (19) we obtain

$$|s' - \hat{s}|_2 > \frac{1}{2}cf(1-2g)^4 > \frac{(1+b_k)}{\sqrt{k}},$$

which means that the alternative tree ($\mathcal{T}_2$ or $\mathcal{T}_3$) is not a good tree. Thus, when event E occurs, the set of good trees consists of $\mathcal{T}_1$ and no-other tree (so that method M will select the true tree, namely $\mathcal{T}_1$). Since event E occurs with probability at least $1-\epsilon$ this verifies that M satisfies condition (i) of Lemma 4.3.

We now establish that M satisfies condition (ii). By Lemma 6.2, the probability that M selects $\mathcal{T}_1$ (the true tree) is at least

$$1/3(1-\exp(-b_k/4))$$

(since the probability that $\mathcal{T}_1$ goes into the 'good' set is at least $1-\exp(-b_k/4)$ and there are at most 3 good trees to select and so is bounded away from 0). Since $b_k \geq b_1$ for all k and given the restriction placed on b_1, it follows that the probability that M selects the true tree is at least δ. This verifies that M satisfies condition (ii) of Lemma 4.3.

Acknowledgement We are indebted to Eva Czabarka for her comments on this paper.

References

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

[2] Bininda-Emonds, O. R. P., Brady, S. G., Kim, J., and Sanderson, M. J. (2001) Scaling of accuracy in extremely large phylogenetic trees. *Pacific Symposium on Biocomputing* **6** 547–558.

[3] Colonius, H., and Schulze, H. H. (1981) Tree structures for proximity data. *British Journal of Mathematical and Statistical Psychology*, **34** 167–180.

[4] Dekker, M. C. H. (1986) *Reconstruction Methods for Derivation Trees*, Master's Thesis, Vrije Universiteit, Amsterdam.

[5] Erdős, P. L., Steel, M. A., Székely, L. A., and Warnow, T. J. (1999) A few logs suffice to build (almost) all trees I, *Random Structures and Algorithms* **14(2)** 153–184.

[6] Evans, W., Kenyon, C., Peres, Y., and Schulman, L. J. (2000) Broadcasting on trees and the Ising model. *Adv. Appl. Prob.* **10** 410–433.

[7] Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52(5)** 696–704.

[8] Lovász, L. (1979) *Combinatorial Problems and Exercises*, Akadémiai Kiadó, Budapest, and North–Holland.

[9] Mason, K. (1996) *On a Matter of Uncertain Parentage*, Honours III Project, University of Canterbury.

[10] Mossel, E. (1998) Recursive reconstruction on periodic trees, *Random Structures Algorithms* **13** 81–97.

[11] Neyman, J. (1971) Molecular studies of evolution: a source of novel statistical problems. in *Statistical Decision Theory and Related Topics*, Gupta, S. S. and J. Yackel (eds), New York, Academic Press, 1–27.

[12] Rokas, A., and Carroll, S. B. (2005) More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol. Biol. Evol.* **22(5)** 1337–1344.

[13] Semple, C., and Steel, M. (2003) *Phylogenetics*, Oxford University Press.

[14] Stamatakis, A.P., Ludwig, T., Meier, H. (2004). A fast program for maximum likelihood-based inference of large phylogenetic tress. *Proceedings of the 2004 ACM symposium on applied computing* 197–201, ACM Press, New York, NY, USA.

[15] Steel, M. (1989) *Distributions on Bicoloured Evolutionary Trees,* PhD thesis, Massey University, Palmerston North, New Zealand.

[16] Steel, M., and Charleston, M. (1995) Five surprising properties of parsimoniously coloured trees. *Bull. Math. Biol.* **57(2)** 367–375.

[17] Steel, M. A., Hendy, M. D., and Penny, D. (1998) Reconstructing phylogenies from nucleotide pattern frequencies - a survey and some new results, *Discrete Applied Mathematics* **88** 367–396.

[18] Steel, M. A., and Székely, L. A. (1999) Inverting random functions *Annals of Combinatorics* **3** 103–113.

[19] Steel, M. A., and Székely, L. A. (2002) Inverting random functions II: explicit bounds for the discrete maximum likelihood estimation, with applications, *SIAM J. Discrete Math.* **15(4)** 562–575.

Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand; and Department of Mathematics, University of South Carolina, Columbia SC, USA

E-mail address: `m.steel@math.canterbury.ac.nz, szekely@math.sc.edu`