Optimal Phylogenetic Reconstruction

Constantinos Daskalakis* Elchanan Mossel† Sébastien Roch‡

September 27, 2005

Abstract

One of the major tasks of evolutionary biology is the reconstruction of phylogenetic trees from molecular data. This problem is of critical importance in almost all areas of biology and has a very clear mathematical formulation. The evolutionary model is given by a Markov chain on the true evolutionary tree. Given samples from this Markov chain at the leaves of the tree, the goal is to reconstruct the evolutionary tree. It is crucial to minimize the number of samples, i.e., the length of genetic sequences, as it is constrained by the underlying biology, the price of sequencing etc.

It is well known that in order to reconstruct a tree on n leaves, sequences of length $\Omega(\log n)$ are needed. It was conjectured by M. Steel that for the CFN evolutionary model, if the mutation probability on all edges of the tree is less than $p^* = (\sqrt{2}-1)/2^{3/2}$ than the tree can be recovered from sequences of length $O(\log n)$. This was proven by the second author in the special case where the tree is "balanced". The second author also proved that if all edges have mutation probability larger than p^* then the length needed is $n^{\Omega(1)}$. This "phase-transition" in the number of samples needed is closely related to the phase transition for the reconstruction problem (or extremality of free measure) studied extensively in statistical physics and probability.

Here we complete the proof of Steel's conjecture and give a reconstruction algorithm using optimal (up to a multiplicative constant) sequence length. Our results further extend to obtain optimal reconstruction algorithm for the Jukes-Cantor model with short edges. All reconstruction algorithms run in time polynomial in the sequence length.

The algorithm and the proofs are based on a novel combination of combinatorial, metric and probabilistic arguments.

Keywords: Phylogenetics, CFN model, Ising model, phase transitions, reconstruction problem, Jukes Cantor.

*Computer Science, U.C. Berkeley, CA 94720. Email: costis@eecs.berkeley.edu. Supported by CIPRES (NSF ITR grant # NSF EF 03-31494).

†Statistics, U.C. Berkeley, CA 94720. Email: mossel@stat.berkeley.edu. Supported by a Miller fellowship in Statistics and Computer Science, by a Sloan fellowship in Mathematics and by NSF grants DMS-0504245 and DMS-0528488

‡Statistics, U.C. Berkeley, CA 94720. Email: sroch@stat.berkeley.edu. Supported by CIPRES (NSF ITR grant # NSF EF 03-31494), FQRNT, NSERC and a Loève Fellowship.

1 Introduction

Phylogenies are used in evolutionary biology to model the stochastic evolution of genetic data on the ancestral tree relating a group of species. The leaves of the tree correspond to (known) extant species. Internal nodes represent extinct species while the root of the tree represents the most recent ancestor to all species in the tree. Following paths from the root to the leaves, each bifurcation indicates a speciation event whereby two new species are created from a parent. We refer the reader to [8] for an excellent introduction to Phylogeny.

The underlying assumption is that genetic information evolves from the root to the leaves according to a Markov model on the tree. This genetic information may consist of DNA sequences, proteins etc. Suppose for example that the genetic data consists of (aligned) DNA sequences and lets follow the evolution of the first letter in all sequences. This collection, named the first *character*, evolves according to Markov transition matrices on the edges. The root is assigned one of the four letters A, C, G and T. Then this letter evolves from parents to descendants according to the Markov matrices on the edges connecting them.

The vector of the i'th letter of all sequences is called the i'th *character*. It is further assumed that the character are i.i.d. random variables. In other words, each site in a DNA sequence is assumed to mutate independently from its neighbors according to the same mutation mechanism. Naturally, this is an over-simplification of the underlying biology. Nonetheless, the model above may be a good model for the evolution of some DNA subsequences and is the most popular evolution model in molecular biology, see e.g. [8]. One of the major tasks in molecular biology, the *reconstruction of phylogenetic trees*, is to infer the topology of the (unknown) tree from the characters (sequences) at the leaves (extant species).

One of the simplest mutation model is given by the Cavender-Farris-Neyman (CFN) model [3, 7, 20]. In this model the character states are 0 and 1 and their a priori probability at the root is $1/2$ each (the 0 and 1 originally correspond to the Purine and Pyrimidine groups). To each edge e there corresponds a mutation parameter $p(e)$ which is the probability that the character mutates along the edge e. In this paper we will be mostly interested in the CFN model.

A problem that is closely related to the Phylogenetic problem is that of inferring the *ancestral state*, i.e., the character state at the root of the tree, given the character at the leaves. This problem was studied earlier in statistical physics, probability and computer science under the title of the *reconstruction problem*, or the "extremality of the free Gibbs measure", see [21, 10, 9]. The reconstruction problem for the CFN model was analyzed in [2, 6, 11, 1, 13].

Roughly speaking, the reconstruction problem is *solvable* when the correlation between the root and the leaves persists no matter how large the tree is. When it is unsolvable, the correlation decays to 0 for large trees. The results of [2, 6, 11, 1, 13] show that for the CFN model, if for all e, it holds that $p(e) \leq p_{\max} < p^*$ then the reconstruction problem is solvable, where

$$p^* = \frac{\sqrt{2}-1}{\sqrt{8}}.$$

If, on the other hand, for all e it holds that $p(e) \geq p_{\min} > p^*$ and the tree is balanced in the sense that all leaves are at the same distance from the root, then the reconstruction problem is unsolvable. Moreover in this case, the correlation between the root state and any function of the character states at the leaves decays as $n^{-\Omega(1)}$.

M. Steel [22] conjectured that when $0 < p_{\min} \leq p(e) \leq p_{\max} < p^*$ for all edges e, one can reconstruct with high probability the phylogenetic tree from $O(\log n)$ characters. Steel's insightful conjecture suggests that there are deep connections between the reconstruction problem and phylogenetic reconstruction.

This conjecture has been proven to hold for trees where all the leaves are at the same distance from the root in [16]. It is also shown there that the number of characters needed when $p(e) \geq p_{\min} > p^*$ for all e is $n^{\Omega(1)}$. The second result intuitively follows from the fact that the topology of the part of the tree that is close to the root is essentially independent of the character at the leaves if the number of characters is not at least $n^{\Omega(1)}$.

The basic intuition behind Steel's conjecture is that since in the regime where $p(e) < p_{\max} < p^*$, there is no decay of the quality of reconstructed sequences, it should be as easy to reconstruct deep trees as it is to reconstruct shallow trees. In [5] (see also [17]) it is shown that "shallow" trees can be reconstructed from $O(\log n)$ characters if all mutation probabilities are bounded away from 0 and $1/2$. The same high-level reasoning has also yielded a complete proof that $O(\log n)$ characters suffice for a homoplasy-free mutation model when all edges are short [19].

Here we give a complete proof of Steel's conjecture. We show that if $0 < p_{\min} \leq p(e) \leq p_{\max} < p^*$ for all edges e of the tree then the tree can be reconstructed from $O(c(p_{\min}, p_{\max}, \delta) \log n)$ characters with error probability at most δ. This result implies that sequences of logarithmic length suffice to reconstruct the tree also for the Jukes-Cantor model when all the edges are sufficiently short.

1.1 Definitions and results

Let T be a tree. Write $\mathcal{V}(T)$ for the nodes of T, $\mathcal{E}(T)$ for the edges of T and $\mathcal{L}(T)$ for the leaves of T. If the tree is rooted, then we denote by $\rho(T)$ the root of T. Unless stated otherwise, all trees are assumed to be *binary* (all internal degrees are 3) and it is further assumed that $\mathcal{L}(T)$ is labeled.

Let T be a tree equipped with a path metric $d : \mathcal{E}(T) \to \mathcal{R}_+$. d will also denote the induced metric on $\mathcal{V}(T)$:

$$d(v, w) = \sum \{d(e) : e \in \mathtt{path}_T(v, w)\}, \tag{1}$$

for all $v, w \in \mathcal{V}(T)$, where $\mathtt{path}_T(x, y)$ is the path (sequence of edges) connecting x to y in T.

We will further assume below that the length of all edges is bounded between f and g for all $e \in E$. In other words, for all $e \in \mathcal{E}(T)$,

$$f \leq d(e) \leq g. \tag{2}$$

We now define the evolution process on a rooted tree equipped with a path metric d. The process is determined by a rooted tree $T = (V, E)$ equipped with a path metric d and a *mutation rate matrix* Q. We will be mostly interested in the case where $Q = \left(\begin{smallmatrix} -1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$ corresponding to the CFN model and in the case where Q is a 4×4 matrix given by $Q_{i,j} = 1 - 4\delta(i = j)$ corresponding to the Jukes-Cantor model. To edge e of length $d(e)$ we associate the mutation matrix $M^e = \exp(d(e)Q)$.

In the mutation model on the tree T rooted at ρ each vertex iteratively chooses its state from the state at its parent by an application of the Markov transition rule M^e. We assume that all edges in E are directed away from the root. Thus the probability distribution on the tree is the probability distribution on $\{0, 1\}^V$ ($\{A, C, G, T\}^V$) is given by

$$\overline{\mu}[\sigma] = \pi(\sigma(\rho)) \prod_{(x \to y) \in E} M^e_{\sigma(x), \sigma(y)}, \tag{3}$$

where π is given by the uniform distribution at the root, so that $\pi(0) = \pi(1) = 1/2$ for the CFN model and $\pi(A) = \pi(C) = \pi(G) = \pi(T) = 1/4$ for the JC model.

We let the measure μ denote the marginal of $\overline{\mu}$ on the set of leaves which we identify with $[n]$. Thus

$$\mu(\sigma) = \sum \{\overline{\mu}(\tau) : \forall i \in [n], \tau(i) = \sigma(i)\}.$$

The measure μ defines the probability distribution at the leaves of the tree.

We note that both for the CFN model and for the JC model, the mutation matrices M^e are in fact very simple. For the CFN model, with probability $p(e) = (1 - \exp(-2d(e)))/2$, there is a mutation and, otherwise, there is no mutation. Similarly for the JC model with probability $p(e) = (1 - \exp(-4d(e)))/4$ each of the three possible mutations occur. In particular writing

$$g^* = \frac{\log 2}{4}, \tag{4}$$

we may formulate the result on the reconstruction problem for the phase transition of the CFN model as follows:

If $d(e) \leq g < g^*$ for all e then the reconstruction problem is solvable.

We will be interested in reconstructing phylogenies in this regime. The objective here is to reconstruct the underlying tree T whose internal nodes are unknown from the collection of sequences at the leaves. Let $\mathcal{T}$ represent the set of all all binary trees on n leaves and $\mathcal{M}$ represent a family of mutation matrices corresponding to edges e whose length d satisfies:

$$0 < f \leq d \leq g < g^*, \tag{5}$$

where g^* is given in (4) and f is an arbitrary positive constant. Let $\mathcal{T} \otimes \mathcal{M}$ denote the set of all phylogenetic trees, where the underlying tree T is in $\mathcal{T}$ and all mutation matrices on the edges are in $\mathcal{M}$. Under mild conditions [4], different elements in $\mathcal{T} \otimes \mathcal{M}$ correspond to different measures μ. Below, we identify the measure μ with the corresponding element of $\mathcal{T} \otimes \mathcal{M}$. We are interested in finding an efficiently computable map Ψ such that $\Psi(\sigma_\partial^1, \ldots, \sigma_\partial^k) \in \mathcal{T}$. Moreover, we require that for every distribution $\mu \in \mathbf{T} \otimes M$ which is defined on a tree T, if $\sigma_\partial^1, \ldots, \sigma_\partial^k$ are generated independently from μ, then with high probability $\Psi(\sigma_\partial^1, \ldots, \sigma_\partial^k) = T$. In [5], it is shown there exists a polynomial time algorithm that reconstructs the topology from $k = \text{poly}(n, 1/\delta)$ characters. Here, we prove the following.

Theorem 1 *Let $f > 0$ and $g < g^*$. Consider the CFN model on binary trees. Then there exists a polynomial time algorithm that reconstructs the topology of the tree from $k = c(f, g, \delta) \log n$ characters with error probability at most δ.*

Corollary 1 *Consider the JC model on binary trees where all edges satisfy*

$$0 < f \leq d(e) \leq g < g^*/2.$$

Then there exists a polynomial time algorithm that reconstructs the topology of the tree from $c(f, g, \delta) \log n$ characters with error probability at most δ.

1.2 Properties of the majority function

In this subsection we quote some of the results we are using from [17]. The results of [17] are stated assuming that the character values are ± 1 instead of $0/1$. Further instead of using the mutation probability $0 \leq p(e) \leq 1/2$ it uses $\theta(e) = 1 - 2p(e)$ which satisfies $0 \leq \theta(e) \leq 1$. Note that in terms of θ we have reconstruction solvability whenever $\theta(e) \geq \theta > \theta_*$ for all e where $2\theta_*^2 = 1$.

For the CFN model both the majority algorithm [10] and recursive majority algorithms [14] are effective in reconstructing the root value (for other models in general, most simple reconstruction algorithms are not effective all the way to the reconstruction threshold [15, 18, 12]).

We now define formally the function Maj. Note that when the number of inputs is even, this function is *randomized*.

Definition 1 *Let* $\text{Maj} : \{-1,1\}^d \to \{-1,1\}$ *be defined as:*

$$\text{Maj}(x_1,\ldots,x_d) = \text{sign}(\sum_{i=1}^{d} x_i + 0.5\omega),$$

where ω *is an unbiased* ± 1 *variable which is independent of the* x_i*. Thus when* d *is odd,*

$$\text{Maj}(x_1,\ldots,x_d) = \text{sign}(\sum_{i=1}^{d} x_i).$$

When d *is even,*

$$\text{Maj}(x_1,\ldots,x_d) = \text{sign}(\sum_{i=1}^{d} x_i),$$

unless $\sum_{i=1}^{d} x_i = 0$*, in which case* $\text{Maj}(x_1,\ldots,x_d)$ *is chosen to be* ± 1 *with probability* $1/2$*.*

Definition 2 *Let* $T = (V,E)$ *be a tree rooted at* ρ *with leaf set* ∂T*. For functions* $\theta' : E \to [0,1]$ *and* $\eta' : \partial T \to [0,1]$*, let* $CFN(\theta',\eta')$ *be the CFN model on* T *where*

- $\theta(e) = \theta'(e)$ *for all* e *which is not adjacent to* ∂T*, and*
- $\theta(e) = \theta'(e)\eta'(v)$ *for all* $e = (u,v)$*, with* $v \in \partial T$*.*

Let

$$\widehat{\text{Maj}}(\theta',\eta') = \mathbf{E}[+\text{Maj}(\sigma_{\partial T})|\sigma_\rho = +1] = \mathbf{E}[-\text{Maj}(\sigma_{\partial T})|\sigma_\rho = -1],$$

where σ *is drawn according to* $CFN(\theta',\eta')$*.*

For functions θ and η as above, we abbreviate by writing $\min\theta$ for $\min_E \theta(e)$, $\max\eta$ for $\max_{v\in\partial T}\eta(v)$, etc. The function $\widehat{\text{Maj}}$ measures how well majority calculates the color at the root of the tree.

Theorem 2 *[17] Let*

$$a(d) = 2^{1-d}\lceil\frac{d}{2}\rceil\binom{d}{\lceil\frac{d}{2}\rceil}. \tag{6}$$

For all ℓ *integer,* $\theta_{\min} \in [0,1]$ *and* $0 \le \alpha < a(b^\ell)\theta_{\min}^\ell$*, there exists* $\beta = \beta(b,\ell,\theta_{\min},\alpha) > 0$ *such that the following hold. Let* T *be an* ℓ*-level balanced* b*-ary tree, and consider the* $CFN(\theta,\eta)$ *model on* T*, where* $\min\theta \ge \theta_{\min}$ *and* $\min\eta \ge \eta_{\min}$*. Then*

$$\widehat{\text{Maj}}(\theta,\eta) \ge \min\{\alpha\eta_{\min},\beta\}. \tag{7}$$

In particular, given b *and* $\theta_{\min}$ *such that* $b\theta_{\min}^2 > h^2 > 1$*, there exist* $\ell(b,\theta_{\min})$*,* $\alpha(b,\theta_{\min}) > h^\ell$ *and* $\beta(b,\theta_{\min}) > 0$*, such that any* $CFN(\theta,\eta)$ *model on the* ℓ*-level* b*-ary tree satisfying* $\min\theta \ge \theta_{\min}$ *and* $\min\eta \ge \eta_{min}$ *must also satisfy (7)*

2 The Algorithm

2.1 Cherry Oracle

At a high level, our reconstruction algorithm proceeds from a simple idea: it builds the tree one level of cherries at a time. In a binary tree, a *cherry* is a pair of leaves at graph distance 2. To see how this would work, imagine that we had access to a "cherry oracle", i.e. a function $C(u, v, T)$ that returns the parent of the pair of leaves $\{u, v\}$ if the latter form a cherry in the tree T (and say 0 otherwise). Then, we could perform the following "cherry picking" algorithm:

- Current tree: $T' := T$;
- Repeat until T' is empty,
  - For all $(u, v) \in \mathcal{L}(T') \times \mathcal{L}(T')$, if $w := C(u, v, T') \neq 0$, set $\mathrm{Parent}(u) := \mathrm{Parent}(v) := w$;
  - Remove from T' all cherries uncovered at the previous step;

Unfortunately, the cherry oracle cannot be simulated from short sequences at the leaves. Indeed, short sequences provide only local metric information on the structure of the tree. For instance, consider a short linear tree attached to the root of a deep complete binary tree. From local metric information, it is impossible to tell which "end" of the linear tree is attached to the complete binary tree.

2.2 Blindfolded Cherry Picking

Nevertheless, the above scheme can be roughly followed by making a simple modification: at every level, pick not only true cherries but also "local" cherries; and add a procedure that cleans up "fake" cherries when more information becomes available. We call this new algorithm, detailed in Figure 1 (see subsequent figures for subroutines), BLINDFOLDED CHERRY PICKING (BCP). A further issue tackled by BCP is that the true sequences at internal nodes are unknown. For this, BCP reconstructs biased estimates of the internal sequences as in [17] and uses these biased sequences to obtain local information deeper inside the tree. The description of the algorithm uses the following notation and conventions:

- $T_{\leq w}^{\mathrm{Child}}$ is the tree made of the children of w as defined by the function Child.
- For sequences $\sigma_u, \sigma_v \in \{\pm 1\}^k$,

$$\widehat{\mathrm{Dist}}\,(\sigma_u, \sigma_v) = -\frac{1}{2}\log\left[\left(\frac{1}{k}\sum_{t=1}^{k}\sigma_u^t\sigma_v^t\right)_+\right].$$

- A *g-cherry* is a cherry where both edges have length less or equal to g.
- Let $M > 0$. Let T be a tree and F be the subforest of T where we keep all the leaves and only those nodes with the following property: they are on a path of length at most M between two leaves of T. We say that a pair of leaves $\{u, v\}$ is an *M-local g-cherry* in T if $\{u, v\}$ is a g-cherry in F.
- ε_2 is a constant to be determined in section 3.
- The variables $i, j, \widehat{L}_i, \widehat{C}_i, \hat{d}_i, \hat{\gamma}, \hat{\sigma}$ are global.
- A *pseudoleaf* is a current active node.

Algorithm BLINDFOLDED CHERRY PICKING (BCP)
Input: samples at the leaves;
Output: estimated topology;

- **Step 0: Initialization**
  - Iteration counter: $i := 0$; Node counter: $j := n$;
  - Active pseudoleaf set: $\widehat{L}_0 := [n]$;
  - Leaf sequences: $\forall i \in [n]$, $\hat{\sigma}_i := \sigma_i$;
- **Step 1: Distance Estimation**
  - For all $(u, v) \in \widehat{L}_i \times \widehat{L}_i$, set $\hat{d}_i(u, v) :=$ DISTEST(u, v);
- **Step 2: Cherry Identification**
  - Parent pseudoleaf set: $\widehat{L}_{i+1} := \widehat{L}_i$;
  - Resolved cherries: $\widehat{C}_i := \emptyset$;
  - For all $(u_0, v_0) \in \widehat{L}_i \times \widehat{L}_i$ such that $u_0 < v_0$, apply CHERRYID (u_0, v_0);
- **Step 3: Sequence Reconstruction**
  - For all $(u, w, v) \in \widehat{C}_i$, set $\hat{\sigma}_w :=$ SEQREC(u, w, v);
- **Step 4: Fake Cherry Detection**
  - For all $(u_0, u_1) \in \widehat{L}_{i+1} \times \widehat{L}_{i+1}$ with $u_0 < u_1$, perform FAKECHERRY(u_0, u_1);
- **Step 5: Termination**
  - If $|\widehat{L}_{i+1}| \leq 3$, compute the length of the missing edges; Output the reconstructed tree.
  - Else, set $i := i + 1$, and go to Step 1.

Figure 1: Algorithm BLINDFOLDED CHERRY PICKING.

Algorithm FOURPOINT
Input: Four nodes and distances between them;
Output: quartet split (if four input nodes) and edge weights;

- Perform four point method to find the right split and estimate the internal edge of the quartet;
- Do at most 4 applications of the four point method to estimate all other edge lengths (using a scheme similar to that in routine DISTEST; see proof of Lemma 7).

Figure 2: Subroutine FOURPOINT.

Algorithm DISTEST
Input: pair of pseudoleaves (u, v); **Output:** estimated distance between u and v;

- If u and v are leaves,
  - Compute $\hat{d}_i(u, v) := \widehat{\mathrm{Dist}}(\hat{\sigma}_u, \hat{\sigma}_v)$;
- If one of u, v is a leaf (say u),
  - Let v', v'' be the children of v;
  - Compute the correlation distances between u, v', v'' and use four point method to deduce the distance between u and v;
- If none of u, v is a leaf,
  - Let u', u'' (resp. v', v'') be the children of u (resp. v);
  - Compute the correlation distances between u', u'', v', v'' and use four point method to deduce the distance between u and v;

Figure 3: Subroutine DISTEST.

Algorithm CHERRYID
Input: pair of pseudoleaves (u_0, v_0);

- $\mathrm{IsCherry} :=$ TRUE;
- *Test 1 [Distance less than $2g + \varepsilon_2$]:* If $\hat{d}_i(u_0, v_0) > 2g + \varepsilon_2$, then $\mathrm{IsCherry} :=$ FALSE;
- *Test 2 [Local cherry]:* Let R_{5g} be the set of all $(u_1, v_1) \in \widehat{L}_i \times \widehat{L}_i$ such that $u_1 < v_1$, $\{u_0, v_0\} \cap \{u_1, v_1\} = \emptyset$, and
$$\max\left\{\hat{d}_i(x_0, x_1) \,:\, x_\iota \in \{u_\iota, v_\iota\}\right\} \leq 5g + \varepsilon_2.$$
Then:
  - If R_{5g} is empty, then $\mathrm{IsCherry} :=$ FALSE;
  - Otherwise, perform FOURPOINT(u_0, v_0, u_1, v_1); If (u_0, v_0) is not a $(g+\varepsilon_2)$-cherry in $\{u_0, v_0, u_1, v_1\}$, then set $\mathrm{IsCherry} :=$ FALSE;
- If $\mathrm{IsCherry} =$ TRUE,
  - Set $j := j + 1$ and $w := j$;
  - Add w to $\widehat{L}_{i+1}$, add (u, w, v) to $\widehat{C}_i$, and remove u, v from $\widehat{L}_{i+1}$; Update parenting relationships;
  - Let $\hat{\gamma}(u_0, w)$ and $\hat{\gamma}(u_1, w)$ be the edge lengths computed above (from one of the "witness" pair).

Figure 4: Subroutine CHERRYID.

Algorithm SEQREC
Input: cherry $(u, w, v) \in \widehat{C}_i$; **Output:** reconstructed sequence at w;

- Let l be the required number of levels from Theorem 2;
- Consider the subtree $T_w^{(l)}$ consisting of all the nodes in $T_{\leq w}^{\mathrm{Child}}$ at topological distance at most l from w;
- Let $L_w^{(l)}$ be the leaf set of $T_w^{(l)}$;
- For each node x in $L_w^{(l)}$,
  - Let $\mathrm{Top}(x)$ be its topological distance from w in $T_w^{(l)}$;
  - Set the *weight of* w to be $h(x) := 2^{l-\mathrm{Top}(x)}$;
- Return $\hat{\sigma}_w := \mathrm{Maj}_h\left(\sigma_x;\ x \in L_w^{(l)}\right)$ (sitewise weighted majority with uniform breaks);

Figure 5: Subroutine SEQREC.

Algorithm FAKECHERRY
Input: pseudoleaves u_0, u_1;

- For $\iota = 0, 1$, set $T_\iota := T_{\leq u_\iota}^{\mathrm{Child}}$ and denote C_ι the set of cherries in T_ι;
- Compute all pairwise distances $\hat{d}$ between T_0 and T_1 using DISTEST (some of these distances are actually wrong);
- $\forall (\kappa_0, \kappa_1) \in C_0 \times C_1$ with $\kappa_\iota = (x_\iota, z_\iota, y_\iota)$, set $\hat{d}_M(\kappa_0, \kappa_1) = \max\{\hat{d}(v_0, v_1) \,:\, v_\iota \in \{x_\iota, y_\iota\}\}$;
- For $\iota = 0, 1$, unless u_ι is not in a $\widehat{C}_i$ or $u_{1-\iota}$ is a leaf, do
  - Let $\kappa_r = (x_r, u_\iota, y_r)$ be the cherry including u_ι;
  - Set $C' := \{\kappa \in C_{1-\iota} \,:\, \hat{d}_M(\kappa_r, \kappa) \leq 25g\}$ (break if empty);
  - Set $\mathrm{Stop} :=$ FALSE;
  - While $C' \neq \emptyset$ and $\mathrm{Stop} =$ FALSE,
    - Let $\kappa = (x, z, y)$ be the lowest cherry in C';
    - *[Collision Test 1]* Let w be the (possibly new) node at the intersection of the triplet $\{x_r, x, y\}$, use the four point method on $\{x_r, x, y\}$ to compute the distance between x and w, say h (using a scheme similar to that in routine DISTEST), check whether $h \neq \hat{\gamma}(x, z)$ (up to $2\varepsilon_2$);
    - *[Collision Test 2]* Perform the previous step again with y_r rather than x_r;
    - If in both tests $h \neq \hat{\gamma}(x, z)$, then set $\mathrm{Stop} :=$ TRUE and set $w_{1-\iota} := w$; otherwise remove κ from C'.
- For $\iota = 0, 1$, perform BUBBLE(w_ι, u_ι).

Figure 6: Subroutine FAKECHERRY.

Algorithm BUBBLE
Input: node x, pseudoleaf w';

- Let (u, y) be the edge on which x is located with $y = \mathrm{Parent}(u)$;
- Add u to $\widehat{L}_{i+1}$;
- Set $z := u$;
- While $z \neq w'$,
  - Add $\mathrm{Sister}(z)$ to $\widehat{L}_{i+1}$;
  - Set $z := \mathrm{Parent}(z)$.
- Remove w' from $\widehat{L}_{i+1}$;

Figure 7: Subroutine BUBBLE.

3 Analysis

In this section, we establish that BCP reconstructs the phylogeny correctly. There are two main technical aspects to the proof. The probabilistic part follows [17]. We focus rather on the combinatorial part where the novelty and complexity of BCP lies. There, we first establish a number of combinatorial properties of the current forest $\widehat{\mathcal{F}}_i$ grown by BCP. We then prove that the "correctly reconstructed subforest" of $\widehat{\mathcal{F}}_i$ increases in size at every iteration.

3.1 Preliminaries

The following notation is used in the proofs.

- T is the phylogenetic tree that produced the data.
- $s = c \log n$ is the number of samples available at the leaves. The constant c is to be determined later.
- $\delta = 1/n^\gamma$ is the probability of error in every application of lemma 3. Since we use a union bound at the end of the argument, we need n^γ to be much bigger than the total number of applications of lemma 3. Thus γ is some large constant independent of n.
- $0 < f < g < +\infty$ are lower and upper bounds on the length of every edge in T.
- $\varepsilon > 0$ is a fixed constant.

In the following discussion, a *subtree* refers to a subgraph of a tree induced by a subset of the nodes. (We sometimes apply this definition to a directed tree, in which case we actually refer to the undirected version of the tree.) We borrow the following notions from [16].

Definition 3 (Edge Disjointness) *Let $\mathtt{path}_T(x, y)$ be the path (sequence of edges) connecting x to y in T. We say that two subtrees T_1, T_2 of T are* edge disjoint *if*

$$\mathtt{path}_T(u_1, v_1) \cap \mathtt{path}_T(u_2, v_2) = \emptyset,$$

for all $u_1, v_1 \in \mathcal{L}(T_1)$ and $u_2, v_2 \in \mathcal{L}(T_2)$. We say that T_1, T_2 are edge sharing *if they are not edge disjoint. (If T_1 and T_2 are directed, we take this definition to refer to their underlying undirected version.)*

Finally, we define the notion of a *collision* between two trees.

Definition 4 (Collisions) *Suppose that T_1 and T_2 are edge disjoint subtrees of T. We say that T_1* and *T_2* collide at distance *d, if the path $\mathtt{path}_T(\rho(T_1), \rho(T_2))$ has non-empty intersection with $\mathcal{E}(T_1) \cup \mathcal{E}(T_2)$ and the length of the shortest path between T_1 and T_2 is at most d.*

In other words, T_1 and T_2 collide at distance d, if the shortest path between T_1 and T_2 is of length at most d and this path does not contain either $\rho(T_1)$ or $\rho(T_2)$.

3.2 Probabilistic Lemmas

Assume that g satisfies the inequality $2e^{-2g} > 1$, which defines the space of values of g for which full reconstruction with $O(\log n)$ samples at the leaves is not forbidden by [17]. Also, fix the constant $\varepsilon < f/2$ such that if $g' = g + \varepsilon$ then g' satisfies $2e^{-2g'} > 1$. The following lemmas are key to our proof.

Definition 5 (Bias) *Suppose v is the root of a tree T. Let Ψ be an antisymmetric function on $\{\pm 1\}^{\mathcal{V}(T)}$ to $\{\pm 1\}$ (i.e, $\Psi(-x) = -\Psi(x)$). Let σ be a character generated by the CFN model on T. Then, the random variable $\tau_v = \sigma(v)\Psi(\sigma)$ is called the* reconstruction bias *of Ψ at v on T.*

Lemma 1 (Reconstruction Bias) *$\exists \varepsilon_1 = \varepsilon_1(g') > 0$ such that if the following hold:*

- *T' is a binary tree rooted at v with edges of length at most g',*
- *$\sigma : \mathcal{V}(T') \to \{\pm 1\}$ is generated according to the CFN model of evolution on T',*

then we can reconstruct a state $\hat{\sigma}(v)$ at the root of T' so that $\mathrm{Corr}(\sigma(v), \hat{\sigma}(v)) \geq \varepsilon_1$. In other words if $\sigma|_{\mathcal{L}} : \mathcal{L}(T') \to \{\pm 1\}$ denotes the value of the character at the leaves, then there exists a randomized function $\Psi : \{\pm 1\}^{\mathcal{L}(T')} \to \{\pm 1\}$ such that it's bias τ_v satisfies $\mathbb{E}[\tau_v] \geq \varepsilon_1$. Moreover, $\mathbb{E}[\tau_v | \sigma(v) = 1] = \mathbb{E}[\tau_v | \sigma(v) = -1]$.

Proof: The proof follows from Theorem 2 by recursively applying the majority function. Let ρ be the root of the tree. Consider the set of all nodes that are either leaves at distance at most ℓ from ρ or internal nodes at distance exactly ℓ from ρ. By induction, we may assume that we have reconstructed the characters at these nodes with correlation at least η. Then the majority of these values (where nodes at distance $r < \ell$ are taken with multiplicity $2^{\ell - r}$) will also give correlation at least η with the original character at the root.

The second claim follows from the fact that the majority function (and therefore all functions we apply) is antisymmetric. ■

Lemma 2 (Distance Estimation 1) *$\forall \gamma > 1, \forall \varepsilon_2 > 0, \exists c_1 = c_1(\gamma, g', \varepsilon_2) > 0$ such that if the following hold. Let*

- *u, v is a pair of nodes,*
- *$\{\hat{\sigma}_u^t\}_{t=1}^k$, $\{\hat{\sigma}_v^t\}_{t=1}^k$ are reconstructed sequences of length $k = c \log n$, $c > c_1$, with the following properties:*
  - *For all t and $w \in \{u, v\}$, $\hat{\sigma}_w^t$ is of the form $\sigma_w^t \tau_w^t$ where σ_w^t is the value generated by the CFN model and τ_w^t (the reconstruction bias) is i.i.d. on $\{\pm 1\}$ with bias at least ε_1 as in lemma 1,*
  - *The variables $\{\tau_u^t\}_{t=1}^k$, $\{\tau_v^t\}_{t=1}^k$ are all independent.*

Then, there is a reconstruction algorithm such that the following hold with probability at least $1 - n^{-\gamma}$:

- *If the $d(u, v) \leq 1001g$ in T then $d(u, v)$ is estimated up to an additive error of ε_2,*
- *If the $d(u, v) > 1001g$ in T then the algorithm outputs an estimated distance for $d(u, v)$ that is $\geq 1000g$.*

Proof: This proof follows from standard concentration inequalities. ■

Lemma 3 (Distance Estimation 2) *$\forall \gamma > 1, \forall \varepsilon_2 > 0, \exists c_1 = c_1(\gamma, g', \varepsilon_2) > 0$ such that if the following hold:*

- *$uv|xy$ is a quartet of width $\leq 1000g$ in T (the width is the maximal distance between any pair in the quartet),*

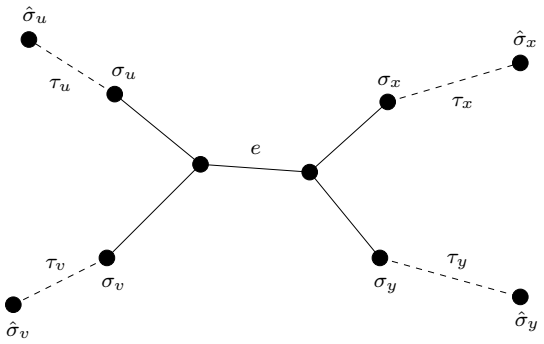

Figure 8: The internal edge is not affected by the bias at the leaves, here represented as extra (dashed) edges.

- *$\{\hat{\sigma}_u^t\}_{t=1}^k$, $\{\hat{\sigma}_v^t\}_{t=1}^k$, $\{\hat{\sigma}_x^t\}_{t=1}^k$, and $\{\hat{\sigma}_y^t\}_{t=1}^k$ are reconstructed sequences of length $k = c \log n$, $c > c_1$, with the following properties:*
  - *For all t and $w \in \{u, v, x, y\}$, $\hat{\sigma}_w^t$ is of the form $\sigma_w^t \tau_w^t$ where σ_w^t is the value generated by the CFN model and τ_w^t (the reconstruction bias) is i.i.d. on $\{\pm 1\}$ with bias at least ε_1 as in lemma 1,*
  - *The variables $\{\tau_u^t\}_{t=1}^k$, $\{\tau_v^t\}_{t=1}^k$ are all independent.*

Then, there is a reconstruction algorithm such that, with probability at least $1 - n^{-\gamma}$, the internal edge of the quartet can be estimated within additive error ε_2.

Proof: Apply the "four point condition" and note that independent biases are equivalent to extra edges in the Markov model. See Figure 8. ■

From here on, we assume that ε_1 is the constant defined by Lemma 1 and the number of samples available at the leaves of the phylogenetic tree is $s = c_1 \log n$, where c_1 is determined by Lemma 3 if we fix $\gamma = 10$ and $\varepsilon_2 < \varepsilon/8$. Our last lemma bounds the error on estimated distances between pseudoleaves. In particular, it accounts for the effect of collisions. We start with a technical observation.

Lemma 4 (Correlation of Antisymmetric Functions) *Let T_1 and T_2 be edge disjoint subtrees of T whose distance is at least αg. For $i = 1, 2$, let $\varphi_i : \{-1, 1\}^{\mathcal{V}(T_i)} \to \{-1, 1\}$ be an antisymmetric function. If σ is a character generated by the CFN model, then*

$$\mathrm{Corr}(\varphi_1(\sigma|_{T_1}), \varphi_2(\sigma|_{T_2})) \leq \exp(-2\alpha g).$$

Proof: We use the random cluster representation of the model. In this representation, an edge e acts as follows:

- with probability $\exp(-2d(e))$ the two endpoints of the edge are identical,
- with probability $1 - \exp(-2d(e))$ the two endpoints are independent.

It is now easy to see that if r is the length of the path connecting the two trees, then with probability $1 - e^{-2r}$ the measures on the two trees are independent. This contributes 0 to the correlation. In the other case, we get a contribution of at most 1. Thus the correlation is bounded by e^{-2r} as needed. ■

This gives immediately:

Lemma 5 (Distance Estimation 3) *Suppose $\mathcal{F} = \{T_1, T_2, \ldots, T_\alpha\}$ is a forest of rooted trees with the following properties:*

- *$\{T_1, T_2, \ldots, T_\alpha\}$ is an edge disjoint subforest of T,*
- *all trees of $\mathcal{F}$ have edges of length at most g',*
- *there is no collision at distance $20g$ in $\mathcal{F}$,*
- *we reconstruct sequences at the roots of the trees in $\mathcal{F}$ using the samples at the leaves of the corresponding tree.*

Then, if we use routine DISTEST *to estimate the distances between every pair of roots of trees in $\mathcal{F}$, the following property is satisfied by the estimated distance $\hat{d}$ with probability at least $1 - n^{-\gamma}$:*

$$\hat{d}(u,v) \leq 12g \vee d(u,v) \leq 12g \Rightarrow |d(u,v) - \hat{d}(u,v)| < \varepsilon_2.$$

3.3 Combinatorial Analysis

The following proposition establishes a number of properties of the forest grown by BCP.

Proposition 1 (Properties of $\widehat{\mathcal{F}}_i$) *The following properties hold at the beginning of* BCP*'s i-th iteration, $\forall i \geq 1$:*

1. [Edge Disjointness] $\widehat{\mathcal{F}}_i = \left\{ T_{\leq u}^{\text{Child}} \,:\, u \in \widehat{L}_i \right\}$ *is an edge disjoint subforest of T.*
2. [Edge Lengths] $\forall u \in \widehat{L}_i$, $T_{\leq u}^{\text{Child}}$ *is a rooted full binary tree with edge lengths at most g'.*
3. [Weight Estimation] *The estimated lengths of the edges in $\widehat{\mathcal{F}}_i$ are within ε_2 from their right values.*
4. [Collisions] *There is no collision at distance $20g$.*

Proof:

$i = 0$: The active set consists of the leaves of T. The claims are therefore trivially true.

$i > 1$: Assume the claims are true at the beginning of the i-th iteration. By doing a step-by-step analysis of the i-th iteration, we show that the claims are still true at the beginning of the $(i+1)$-st iteration. The following lemma follows from Lemma 5.

Lemma 6 (Correctness of DISTEST) *After the completion of step 1, for all $u, v \in \widehat{L}_i$:*

$$\hat{d}_i(u,v) \leq 12g \vee d(u,v) \leq 12g \Rightarrow |d(u,v) - \hat{d}_i(u,v)| < \varepsilon_2.$$

Proof: From the induction hypothesis (Claim 4), it follows that in the beginning of the i-th iteration there is no collision at distance $20g$. So the claim follows from Lemma 5. (A small detail to note is that the sequences at the nodes of the forest were reconstructed in different steps of the algorithm. However, the subtrees that were used for the reconstruction of each node are exactly those in the statement of Lemma 5.) ■

Next, we analyze the routine CHERRYID.

Lemma 7 (Correctness of CHERRYID) *Let u, v be the input pair of pseudoleaves to* CHERRYID. *Let $T' = T - \widehat{\mathcal{F}}_i$ (keeping the nodes in $\widehat{L}_i$) at the beginning of the i-th iteration. Then we have the following.*

- *If $\{u, v\}$ is a $5g$-local g-cherry in T', then it passes all screening tests in* CHERRYID.
- *If $\{u, v\}$ is not a $(5g + 2\varepsilon_2)$-local $(g + 2\varepsilon_2)$-cherry in T', then it is rejected by at least one of the tests in* CHERRYID.

Proof: This result is implied by the following claim. Every time FOURPOINT is called by CHERRYID, say on the four nodes u, v, u', v' where $\{u, v\}$ is the candidate cherry and $\{u', v'\}$ is the witness, then the following hold.

- The trees rooted at u, v, u', v' do not collide.
- The split returned by FOURPOINT is the correct split of the nodes.
- All edge weights of the quartet joining u, v, u', v' are estimated within ε_2 of their correct value.

(This holds also when $u' = v'$.) We now prove this claim.

The subroutine FOURPOINT is called by CHERRYID when the following assumptions are satisfied.

- $\hat{d}_i(u, v) \leq 2g + \varepsilon_2$,
- $\max\left\{\hat{d}_i(u, u'), \hat{d}_i(u, v'), \hat{d}_i(v, u'), \hat{d}_i(v, v')\right\} \leq 5g + \varepsilon_2$.

From Lemma 6, it follows that the above estimated distances are within ε_2 of their correct values. An application of the triangle inequality gives $d(u', v') < 11g$ so that $|\hat{d}_i(u', v') - d(u', v')| < \varepsilon_2$ as well. In fact, all pairwise distances of nodes in the set $\{u, v, u', v'\}$ are smaller than $11g$. Hence, by the induction hypothesis (Claim 4), the four trees rooted at u, v, u', v' do not collide. Therefore, from Lemma 3 and the fact that the quartet joining u, v, u', v' has width at most $11g$, the split of nodes u, v, u', v' is found correctly by the four point method and the length of the internal edge of the quartet is estimated within ε_2 of its correct value.

It remains to show that all other edges of the quartet are estimated within ε_2 of their correct value. Above, we have established that the quartet split computed for the nodes u, v, u', v' is correct. Also, by the induction hypothesis (Claim 1) the trees rooted at u, v, u', v' are edge disjoint subtrees of T. Suppose the quartet joining u, v, u', v' is as depicted in Figure 9 where we are estimating $d(u, z)$. Without loss of generality, assume the algorithm applies the four point method to the set of nodes $\{u_1, u_2, v, v'\}$. It is easy to see that every pair of nodes in the set $\{u_1, u_2, v, v'\}$ has distance $< 7g$ and so the width of the quartet is $< 7g$. Thus, Lemma 3 can be applied and the internal edge of the quartet, i.e. (u, z), is estimated within ε_2 of its correct value.

A similar argument applies to the case $u' = v'$. ■

We are now in a position to prove claims 1, 2, and 3.

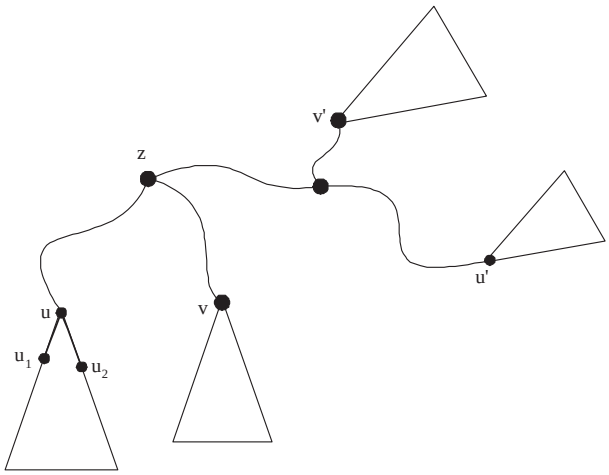

Figure 9: Estimating distance $d(u, z)$.

Lemma 8 (Claims 1, 2, and 3) *At the beginning of the $(i+1)$-st iteration, claims 1, 2, and 3 of the induction hypothesis hold.*

Proof: Since FAKECHERRY only removes edges from the current forest, it is enough to prove that after the completion of Step 3 the resulting forest satisfies claims 1, 2, and 3.

Claim 1. Suppose the resulting forest is not edge disjoint. Also, suppose that, along the execution of Step 2, the forest stopped being edge disjoint when cherry (x, z, y) was added to $\widehat{C}_i$. Then one of the following must be true:

1. There is a pseudoleaf $z' \in \widehat{L}_i \cap \widehat{L}_{i+1}$ (where $\widehat{L}_{i+1}$ is taken at the end of iteration i) such that $\texttt{path}_T(x, y)$ is edge sharing with $T^{\text{Child}}_{\leq z'}$. But then it is not hard to see that there is a collision in $\left\{T^{\text{Child}}_{\leq u} : u \in \widehat{L}_i\right\}$ at distance $3g$ which contradicts the induction hypothesis (Claim 4).

2. There is a pseudoleaf $z' \in \widehat{L}_{i+1} \setminus \widehat{L}_i$ such that $\texttt{path}_T(x, y)$ is edge sharing with $T^{\text{Child}}_{\leq z'}$. We can distinguish the following subcases.

   - $(x', z', y') \in \widehat{C}_i$ and $\texttt{path}_T(x, y)$ is edge sharing with $\texttt{path}_T(x', y')$: in this case $xy|x'y'$ is not the correct split and, by Lemma 7, it is not hard to see that CHERRYID rejects $\{x, y\}$ when performing Test 2. (Note that because both $\{x, y\}$ and $\{x', y'\}$ pass Test 1, and $\texttt{path}_T(x, y)$ and $\texttt{path}_T(x', y')$ are edge-sharing, it follows that $\{x', y'\}$ serves as a "witness" to $\{x, y\}$ in this case.)
   - Otherwise, it is not hard to see that there is a collision at distance $3g$ in $\left\{T^{\text{Child}}_{\leq u} : u \in \widehat{L}_i\right\}$, which contradicts the induction hypothesis (Claim 4).

Claim 2. Follows directly from the description of the algorithm: a cherry (u, x, v) is added to $\widehat{C}_i$ only if $d(u, x)$ and $d(v, x)$ are estimated to be at most $g + \varepsilon_2$, so that the true edge lengths are less than g' by Lemma 6 and the choice of ε_2.

Claim 3. This follows from Lemma 7. ■

It remains to prove Claim 4. This follows immediately from the following analysis of FAKECHERRY.

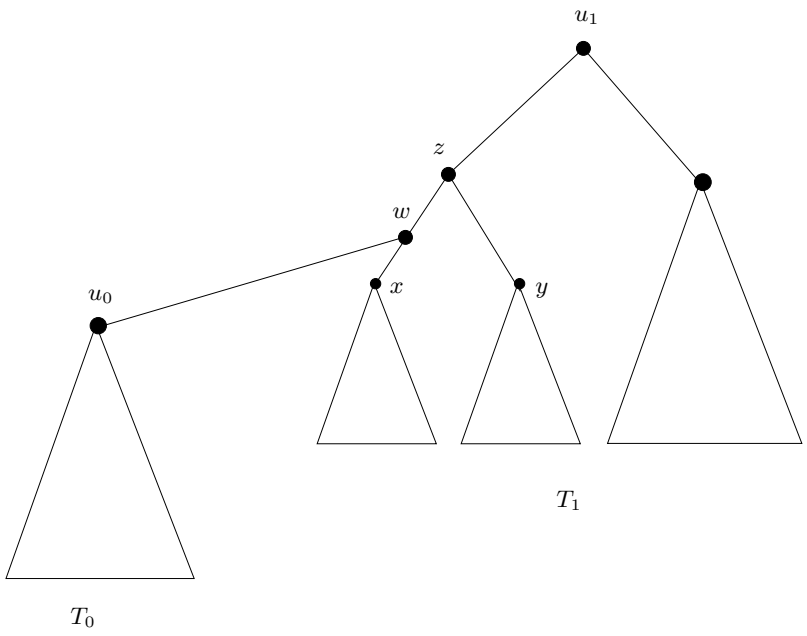

Figure 10: Illustration of routine FAKECHERRY.

Lemma 9 (Collision Removal) *Let $u_0, u_1 \in \widehat{L}_{i+1}$. Suppose $T^{\mathrm{Child}}_{\leq u_0}$ and $T^{\mathrm{Child}}_{\leq u_1}$ collide at distance $20g$. Then* FAKECHERRY *finds the collision.*

Proof: From Claim 4 of the induction hypothesis, it follows that at least one of u_0 or u_1, say u_0 without loss of generality, is such that $(x_r, u_0, y_r) \in \widehat{C}_i$ for some x_r, y_r, and that moreover the path between $T_0 = T^{\mathrm{Child}}_{\leq u_0}$ and $T_1 = T^{\mathrm{Child}}_{\leq u_1}$ starts on (x_r, u_0) or (y_r, u_0). Suppose that e is the edge of T_1 where the collision is located. Consider the set

$$A_{0\to 1} = \{v \in \mathcal{V}(T_1) \ : \ \text{the subtree of } T_1 \text{ rooted at } v \text{ does not contain edge } e\}$$

It is not hard to see that for all $v \in A_{0\to 1}$ the reconstructed sequence at node v is positively correlated with the true sequence and the bias is independent of the biases of the reconstructed sequences at x_r and y_r. Thus, from Lemma 3 and Claim 2 of the induction hypothesis, it follows that $\forall v \in A_{0\to 1} : d(u_0, v) \leq 25g \Rightarrow |\hat{d}_{i+1}(u_0, v) - d(u_0, v)| < \varepsilon_2$. Call $A'_{0\to 1} \subseteq A_{0\to 1}$ the set that contains the nodes $v \in A_{0\to 1}$ such that $\hat{d}_{i+1}(u_0, v) \leq 25g$. Since the collision is at distance $20g$ it follows that $A'_{0\to 1}$ is nonempty and in fact contains at least the lower endpoint of edge e and its sibling in T_1. The routine FAKECHERRY scans the cherries in T_1 starting from the lowest cherry and going up. Therefore, it is easy to see that FAKECHERRY only considers nodes in $A'_{0\to 1}$ and that, by the proof of Lemma 7 (correctness of weight estimations), it stops when it reaches the lower endpoint of e and its sibling. See Figure 10 for an illustration. ■

This concludes the proof of Proposition 1. ■

We now show that, in a precise sense, the algorithm makes progress at every iteration. For this, we consider the following definitions.

Definition 6 (Fixed Subforest) *Let $\mathcal{F}$ be a rooted directed edge disjoint subforest of T with implicit descendance relationship* Child*. Let $u \in \mathcal{V}(\mathcal{F})$. We say that u is* fixed *if $T^{\mathrm{Child}}_{\leq u}$ is fully reconstructed (or*

in other words, $T^{\mathrm{Child}}_{\leq u}$ can be obtained from T by removing (at most) one edge adjacent to u). Note that descendants of a fixed node are fixed themselves. We denote $\mathcal{F}^*$ the (directed) subforest of $\mathcal{F}$ made of all fixed nodes of $\mathcal{F}$. We say that $\mathcal{F}^*$ is the maximal fixed subforest *of $\mathcal{F}$.*

Definition 7 (Bundle) *A* bundle *is a group of four leaves such that:*

- *Any two leaves are at topological distance at most 5;*
- *It includes at least one cherry.*

Proposition 2 (Progress) *Let*

$$\widehat{\mathcal{F}}_i = \left\{ T^{\mathrm{Child}}_{\leq u} \,:\, u \in \widehat{L}_i \right\}$$

(where $\widehat{L}_i$ is taken at the beginning of iteration i) for all $i \geq 0$ with corresponding maximal fixed subforest $\widehat{\mathcal{F}}^*_i$. Then for all $i \geq 0$ (before the termination step), $\widehat{\mathcal{F}}^*_i \subseteq \widehat{\mathcal{F}}^*_{i+1}$ and $|\mathcal{V}(\widehat{\mathcal{F}}^*_{i+1})| > |\mathcal{V}(\widehat{\mathcal{F}}^*_i)|$.

Proof: We first argue that $\widehat{\mathcal{F}}^*_i \subseteq \widehat{\mathcal{F}}^*_{i+1}$. Note that the only routine that removes edges is BUBBLE when called by FAKECHERRY. Because $\widehat{\mathcal{F}}^*_i$ is fully reconstructed, it suffices to show that collisions identified by FAKECHERRY are actual collisions or lie "above" an actual collision—i.e. are on a cherry located on the path between the actual collision and the root. Indeed, since BUBBLE removes only edges "above" presumed collisions, this would then imply that no edge in $\widehat{\mathcal{F}}^*_i$ can be removed. We now prove the claim by analyzing the behavior of FAKECHERRY. We use the notation defined in the routine. Consider the collision tests in FAKECHERRY. The key point is to observe the following:

- if κ is in $\widehat{\mathcal{F}}^*_i$, then at least one of x_r or y_r has a reconstruction bias that is independent from the bias at both x and y; therefore this "correct" witness cannot find a collision (using Lemma 2 and the fact that $\hat{d}_M(\kappa_r, \kappa) \leq 25g$);
- if κ is not in $\widehat{\mathcal{F}}^*_i$, all the cherries above κ (on the path to $u_{1-\iota}$) cannot be in $\widehat{\mathcal{F}}^*_i$ and therefore applying BUBBLE to κ does not modify $\widehat{\mathcal{F}}^*_i$.

This proves the first claim.

For the second claim, assume $\widehat{\mathcal{F}}_i = \{T_1, \ldots, T_\alpha\}$ and $F' \equiv T - \widehat{\mathcal{F}}_i = \{T'_1, \ldots, T'_\beta\}$. F' is the forest obtained from T by removing all the edges in the union of the trees $T_1, \ldots, T_\alpha$. The nodes of F' are all the endpoints of the remaining edges. Since the trees $T_1, \ldots, T_\alpha$ are edge disjoint, the set F' is in fact a subforest of T.

Each leaf v in F' satisfies exactly one of the following:

- **Collision Node:** v a leaf of F' that belongs to a path connecting two vertices in $T_a \in \widehat{F}_i$ but is not the root of T_a (it lies in the "middle" of an edge of T_a).
- **Fixed Pseudoleaf:** v is a root of a fully reconstructed tree $T_a \in \widehat{\mathcal{F}}_i$ (i.e. T_a is also in $\widehat{\mathcal{F}}^*_i$);
- **Colliding Pseudoleaf:** v is a root of a tree $T_a \in \widehat{\mathcal{F}}_i$ that is not in $\widehat{\mathcal{F}}^*_i$ (the tree T_a contains a collision).

A *fixed bundle* is a bundle in F' whose leaves are fixed pseudoleaves. We now prove that F' contains at least one fixed bundle. This immediately implies the second claim. Indeed, it is not hard to see that the cherry in the fixed bundle is found by CHERRYID during the $(i+1)$-st iteration.

Lemma 10 (Fixed Bundle) *Assume Proposition 1 holds at the end of the i-th iteration and let F' as above have at least two internal nodes. Then, F' contains at least one fixed bundle.*

Proof: We first make a few easy observations:

1. A tree with 4 or more leaves contains at least one bundle. (To see this: merge all cherries into leaves; repeat at most twice.)

2. Because of Claim 4 in Proposition 1, collision nodes are at distance at least $20g$ from any other leaf in F'. Therefore, if a tree in F' contains a collision, then it has $\gg 4$ nodes and, from the previous observation, it contains at least one bundle. Moreover, this bundle cannot contain a collision node (since in a bundle all leaves are close).

3. From the previous observations, we get the following: if a tree in F' contains a collision, then either it has a fixed bundle, or it has at least one colliding pseudoleaf.

It is then easy to conclude. Assume there is no collision node in F'. Then, there cannot be any colliding pseudoleaf either and it is easy to see that F' is actually composed of a single tree all of which leaves are fixed. Then there is a fixed bundle by Observation 1 above.

Assume on the contrary that there is a collision node. Let T'_b be a tree in F' with such a node. Then by Observation 3, T'_b either has a fixed bundle, in which case we are done, or it has a colliding pseudoleaf, say v. In the latter case, let T_a be the tree in $\widehat{F}_i$ whose root is v. The tree T_a contains at least one collision node which it shares with a tree in F', say $T'_{b'}$. Repeat the argument above on $T'_{b'}$, and so on.

Note that in each step we "exit" a tree $T_c \in \widehat{F}_i$ via a node that is not the root of $T_c \in \widehat{F}_i$ and enter a new tree $T_d \in \widehat{F}_i$ at its root. Since T is a tree, this process cannot continue forever, and we eventually find a fixed bundle. ■

■

Proof of Theorem 1: By Proposition 1, the current forest is correctly reconstructed. By Proposition 2, after $O(n)$ iterations, there remains at most three nodes in $\widehat{L}_i$. It is easy to see that the termination step correctly reconstructs any missing edge. So when the BCP algorithm terminates, it outputs the tree T (as an undirected tree) with high probability and all estimated edges are within ε_2 of their correct value. ■

Acknowledgments

S.R. thanks Martin Nowak and the Program for Evolutionary Dynamics at Harvard University where part of this work was done. C. D. and E. M. thank Satish Rao for interesting discussions. E.M. thanks M. Steel for his enthusiastic encouragement for studying the connections between the reconstruction problem and phylogeny.

References

[1] N. Berger, C. Kenyon, E. Mossel, and Y. Peres. Glauber dynamics on trees and hyperbolic graphs. *Probab. Theory Related Fields*, 131(3):311–340, 2005. Extended abstract by Kenyon, Mossel and Peres appeared in proceedings of 42nd IEEE Symposium on Foundations of Computer Science (FOCS) 2001, 568–578.

[2] P. M. Bleher, J. Ruiz, and V. A. Zagrebnov. On the purity of the limiting Gibbs state for the Ising model on the Bethe lattice. *J. Statist. Phys.*, 79(1-2):473–482, 1995.

[3] J. A. Cavender. Taxonomy with confidence. *Math. Biosci.*, 40(3-4), 1978.

[4] J. Chang. Full reconstruction of markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.*, 137(51–73), 1996.

[5] P. L. Erdös, M. A. Steel;, L. A. Székely, and T. A. Warnow. A few logs suffice to build (almost) all trees (part 1). *Random Structures Algorithms*, 14(2):153–184, 1999.

[6] W. S. Evans, C. Kenyon, Yuval Y. Peres, and L. J. Schulman. Broadcasting on trees and the Ising model. *Ann. Appl. Probab.*, 10(2):410–433, 2000.

[7] J. S. Farris. A probability model for inferring evolutionary trees. *Syst. Zool.*, 22(4):250–256, 1973.

[8] J. Felsenstein. *Inferring Phylogenies*. Sinauer, New York, New York, 2004.

[9] H. O. Georgii. *Gibbs measures and phase transitions*, volume 9 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, 1988.

[10] Y. Higuchi. Remarks on the limiting Gibbs states on a $(d+1)$-tree. *Publ. Res. Inst. Math. Sci.*, 13(2):335–348, 1977.

[11] D. Ioffe. On the extremality of the disordered state for the Ising model on the Bethe lattice. *Lett. Math. Phys.*, 37(2):137–143, 1996.

[12] S. Janson and E. Mossel. Robust reconstruction on trees is determined by the second eigenvalue. *Ann. Probab.*, 32:2630–2649, 2004.

[13] F. Martinelli, Alistair A. Sinclair, and D. Weitz. Glauber dynamics on trees: boundary conditions and mixing time. *Comm. Math. Phys.*, 250(2):301–334, 2004.

[14] E. Mossel. Recursive reconstruction on periodic trees. *Random Structures Algorithms*, 13(1):81–97, 1998.

[15] E. Mossel. Reconstruction on trees: beating the second eigenvalue. *Ann. Appl. Probab.*, 11(1):285–300, 2001.

[16] E. Mossel. Phase transitions in phylogeny. *Trans. Amer. Math. Soc.*, 356(6):2379–2404, 2004.

[17] E. Mossel. Distorted metrics on trees and phylogenetic forests. Submitted. Availible at Arxiv: math.CO/0403508, 2005.

[18] E. Mossel and Y. Peres. Information flow on trees. *Ann. Appl. Probab.*, 13(3):817–844, 2003.

[19] E. Mossel and M. Steel. A phase transition for a random cluster model on phylogenetic trees. *Math. Biosci.*, 187(2):189–203, 2004.

[20] J. Neyman. Molecular studies of evolution: a source of novel statistical problems. In S. S. Gupta and J. Yackel, editors, *Statistical desicion theory and related topics*, pages 1–27. 1971.

[21] F. Spitzer. Markov random fields on an infinite tree. *Ann. Probability*, 3(3):387–398, 1975.

[22] M. Steel. My Favourite Conjecture. http://www.math.canterbury.ac.nz/ mathmas/conjecture.pdf, 2001.