Shrinkage Effect in Ancestral Maximum Likelihood

Elchanan Mossel, Sebastien Roch, and Mike Steel

Abstract—Ancestral maximum likelihood (AML) is a method that simultaneously reconstructs a phylogenetic tree and ancestral sequences from extant data (sequences at the leaves). The tree and ancestral sequences maximize the probability of observing the given data under a Markov model of sequence evolution, in which branch lengths are also optimized but constrained to take the same value on any edge across all sequence sites. AML differs from the more usual form of maximum likelihood (ML) in phylogenetics because ML averages over all possible ancestral sequences. ML has long been known to be statistically consistent—that is, it converges on the correct tree with probability approaching 1 as the sequence length grows. However, the statistical consistency of AML has not been formally determined, despite informal remarks in a literature that dates back 20 years. In this short note, we prove a general result that implies that AML is statistically inconsistent. In particular, we show that AML can "shrink" short edges in a tree, resulting in a tree that has no internal resolution as the sequence length grows. Our results apply to any number of taxa.

Index Terms—Phylogenetic reconstruction, ancestral maximum likelihood, statistical consistency.

1 Introduction

Markov models of site substitution in DNA are the basis for most methods for inferring phylogenies (evolutionary trees) from aligned sequence data. The usual approach is maximum likelihood (ML), which seeks the tree and branch lengths that maximizes the probability of generating the observed data under a Markov process. In the simplest setting, one assumes that sites evolve independently and identically and that the extant sequences (data) label the leaves of the tree—for background on phylogenetics, ML, and other important reconstruction techniques such as maximum parsimony (MP) and neighbor-joining (NJ), see [10]. ML is computationally complicated, and even the problem of finding the optimal branch lengths exactly on a fixed tree has unknown complexity. In ML, one considers all possible ancestral sequences that could have existed within the tree and averages each such "scenario" by its probability. An alternative is to simply consider a single choice of ancestral sequences that has the highest probability—this is a variant of ML that was introduced in 1987 by Barry and Hartigan [3] under the name "most parsimonious likelihood," which later was renamed *ancestral maximum likelihood* (AML) (see, e.g., [1]). The computational complexity of AML is slightly easier than ML, in that given the tree and either the optimal branch lengths or the optimal ancestral sequences, the other "unknown" (ancestral sequences or branch lengths) is readily determined (see, e.g., [2] and [15]). The method can be viewed as being, in some sense, intermediate between ML and MP, which seeks the tree and ancestral sequences that minimizes the total number of site substitutions required to describe the data. Indeed, AML would select the same trees as MP if one further constrained AML so that each edge had the same branch length, as shown in [11].

The recent interest in AML has sprung from computational complexity considerations. First, AML provided a route by which to show that the problem of reconstructing an ML tree from sequences is NP-hard [1], [6]. It turned out that the NP-hardness of ML can be established directly, without invoking AML [17]; however, the relative computational simplicity of AML over ML suggests that it may provide an alternative strategy for reconstructing large trees.

Nevertheless, it is important to know whether the desirable statistical properties of ML carry over to methods such as AML. In particular, ML has long been known to be statistically consistent as a way of estimating tree topologies (see, e.g., [10] and references therein)—that is, as the sequence length grows, the probability that ML will reconstruct the tree that generated the sequences tends to 1. It has also been known (since 1978) that earlier methods such as MP can be statistically inconsistent [9].

However, the statistical consistency of AML is unclear, since the standard Wald-style conditions required to prove consistency (in particular, a fixed parameter space that does not grow with the size of the data) does not apply. Thus, one may suspect that AML might be inconsistent, and indeed, remarks in the literature have suggested this could be the case (see [4] and [12]). However, the absence of a sufficient condition to prove consistency does not constitute proof of inconsistency, and the purpose of this short note is to formally show that AML is statistically inconsistent. More precisely, we show that AML tends to "shrink" short edges in a tree, and this can result in the collapse of the interior edges (and any short pendant edges) to produce a star tree.

The results in this paper rely on probability arguments, based on expansions of the entropy function, and

- *E. Mossel is with the Department of Statistics, University of California, Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860 and the Department of Mathematics and Statistics, Weizmann Institute of Science, Rehovot 76100 Israel. E-mail: mossel@stat.berkeley.edu.*
- *S. Roch is with Microsoft Research, One Memorial Drive, Cambridge, MA 02142. E-mail: Sebastien.Roch@microsoft.com.*
- *M. Steel is with the Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand. E-mail: m.steel@math.canterbury.ac.nz.*

Manuscript received 7 Feb. 2008; revised 27 May 2008; accepted 2 Oct. 2008; published online 14 Oct. 2008.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBBSI-2008-02-0025.
Digital Object Identifier no. 10.1109/TCBB.2008.107.

combinatorial properties of minimal sets of edges that separate each pair of leaves in a tree.

1.1 Problem Statement

1.1.1 CFN Model

We define $[n] = \{0, \ldots, n-1\}$, and we deal with the *Cavender-Farris-Neyman (CFN) model* [5], [8], [14].

Definition 1 (CFN model). *We are given a tree $T = (V, E)$ on n leaves labeled $[n]$ and an assignment of edge probabilities $\mathbf{p} : E \to (0, 1/2)$. A realization of the model is obtained as follows: choose any vertex as a root, pick a state for the root uniformly at random in {0, 1}, and moving away from the root, each edge e flips the state of its ancestor with probability p_e. We denote by X the (random) state at the leaves obtained in this manner. We write $X \sim \mathrm{CFN}(T, \mathbf{p})$. Note that we do not allow substitution probabilities 0 and $1/2$ in order to guarantee that the model is fully identifiable.*

1.1.2 Ancestral Maximum Likelihood

We consider two equivalent formulations of the *AML problem*.

Definition 2 (AML version 1). *The* AML *problem can be stated as follows: given a set S of n binary sequences of length k, find a tree $T = (V, E)$ on n leaves, an assignment $\mathbf{p} : E \to [0, 1]$ of edge probabilities, and an assignment of sequences $\boldsymbol{\lambda} : V \to \{0, 1\}^k$ to the vertices such that*

1. *the sequences at the leaves under $\boldsymbol{\lambda}$ are exactly the sequences from S and*
2. *the quantity*

$$\mathcal{L}(T, \mathbf{p} \mid \boldsymbol{\lambda}) = -\log_2 \left(\prod_{e \in E} p_e^{d_e} (1 - p_e)^{k - d_e} \right) \tag{1}$$

is minimized, where

$$d_{u,v} = \|\lambda_u - \lambda_v\|_1.$$

Note that $\mathcal{L}(T, \mathbf{p}|\boldsymbol{\lambda})$ above is the log-likelihood score under the CFN model with parameters T and $\mathbf{p}$.

The second version, due to [1], is obtained by setting

$$p_e = \frac{d_e}{k}, \tag{2}$$

for all e in the first version.

Definition 3 (AML version 2 [1]). *The* AML *problem can alternatively be stated as follows: given a set S of n binary sequences of length k, find a tree T on n leaves and an assignment of sequences $\boldsymbol{\lambda} : V \to \{0, 1\}^k$ to the vertices such that*

1. *the sequences at the leaves under $\boldsymbol{\lambda}$ are exactly the sequences from S and*
2. *the quantity*

$$\mathcal{H}(T \mid \boldsymbol{\lambda}) = \sum_{e \in E} H\left(\frac{d_e}{k}\right)$$

is minimized; recall that the entropy function is

$$H(p) = -p \log_2 p - (1 - p) \log_2 (1 - p),$$

for $0 \leq p \leq 1$.

Note that in Definition 2 (and implicitly in Definition 3), we allow substitution probabilities exceeding $1/2$, while such values are excluded in our definition of the CFN model. This is in line with a practical optimization perspective where it may be natural to allow all values $0 \leq p_e \leq 1$ and $0 \leq d_e \leq k$. However, our results are also valid when p_e is restricted between 0 and $1/2$ because this constraint does not play a major role in our proof. Moreover, we explicitly allow values $p_e = 0, 1$ as otherwise, the infimum in (1) may not be attained.

1.1.3 Consistency

We denote by $\mathcal{B}_n^{(k)}$ the set of all data sets on n leaves, where sequences have length k. We generally denote a particular data set by $\boldsymbol{\mu}$, where $\boldsymbol{\mu}(i)$ is the sequence at leaf $i \in [n]$. For each data set, a phylogeny estimator $\Phi_n^{(k)}$ returns a tree on n labeled leaves. For modeling purposes, we consider a particular type of data set: if $\mathbb{X} = \{X_1, X_2, \ldots\}$ is a sequence of (possibly random) characters on n leaves, then $\boldsymbol{\mu}_{\mathbb{X}}^{(k)}$ is the data set corresponding to the first k characters.

More formally, a *phylogeny estimator* $\Phi = \{(\Phi_n^{(k)})_{n,k \geq 1}\}$ is a collection of mappings from sequences to trees, that is

$$\Phi_n^{(k)} : \mathcal{B}_n^{(k)} \to \mathcal{T}_n,$$

where $\mathcal{B}_n^{(k)}$ is the set of all assignments of the form

$$\mathcal{B}_n^{(k)} = \left\{ \boldsymbol{\mu} \mid \boldsymbol{\mu} \ : \ [n] \to \{0, 1\}^k \right\}$$

(in other words, $\mathcal{B}_n^{(k)}$ is the set of all {0, 1}-data matrices with n rows and k columns), and $\mathcal{T}_n$ is the set of all trees on n leaves labeled by $[n]$ (without edge parameters). Let $\mathbb{X} = \{X_1, X_2, \ldots\}$ with $X_j : [n] \to \{0, 1\}$. For all $k \geq 1$, we denote by $\boldsymbol{\mu} = \boldsymbol{\mu}_{\mathbb{X}}^{(k)}$ the data set in $\mathcal{B}_n^{(k)}$ such that $(\mu_v)_j = (X_j)_v$ for all $v \in [n]$ and $j = 1, \ldots, k$ (in other words, the data matrix $\boldsymbol{\mu}_{\mathbb{X}}^{(k)}$ has k columns, where X_j is the jth column).

Definition 4 (consistency). *A phylogeny estimator Φ is said to be (statistically)* consistent *if for all n, all trees $T = (V, E) \in \mathcal{T}_n$ and all edge probability assignments $\mathbf{p} : E \to (0, 1/2)$, it holds that*

$$\Phi_n^{(k)} \left(\boldsymbol{\mu}_{\mathbb{X}}^{(k)} \right) \to T$$

almost surely as $k \to +\infty$, where $\mathbb{X} = \{X_1, X_2, \ldots\}$ with $X_1, X_2, \ldots$ independently generated by $\mathrm{CFN}(T, \mathbf{p})$.

1.2 Main Result

Let Φ_{AML} be the *AML phylogeny estimator* for AML version 1, where all edges e with $p_e = 0$ have been contracted. (In cases where several trees can produce an optimal score for the same data set, pick one such optimal tree arbitrarily.)

Theorem 1 (branch shrinkage in AML). *For all $n \geq 3$ and all tree $T = (V, E) \in \mathcal{T}_n$, there is a constant $\beta > 0$, and an edge parameter set of the form $\mathcal{Q}_T = \prod_{e \in E} I_e$ (Cartesian product), where $I_e \subseteq (0, 1/2)$ is an interval of length at least β, such that if $\mathbf{p} \in \mathcal{Q}_T$, then Φ_{AML} returns a star rooted at 0 in the limit $k \to +\infty$ on the data set $\mathbb{X} = \{X_1, \ldots X_k\}$ with $X_1, \ldots, X_k$ independently generated by $\mathrm{CFN}(T, \mathbf{p})$.*

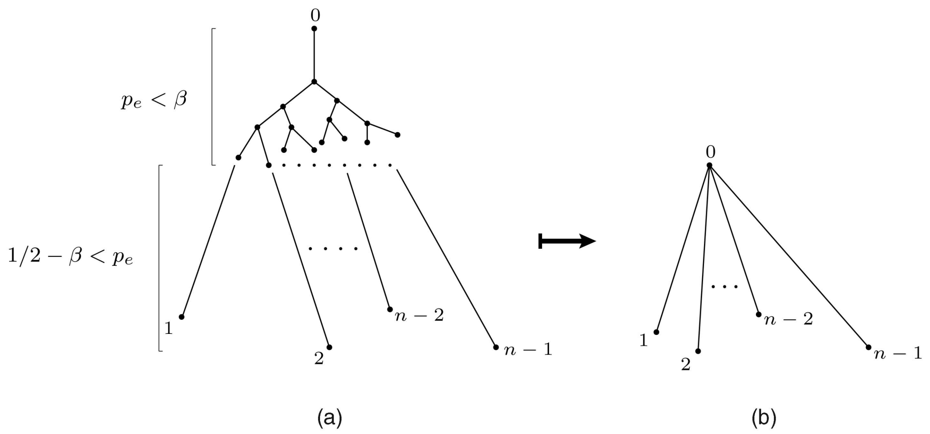


Fig. 1. The shrinkage effect. For the tree on (a), AML will reconstruct the star tree (b) from sufficiently long sequences.

The phenomenon described in Theorem 1 is illustrated in Fig. 1. We refer to $\mathcal{Q}_T$ informally as a "shrinkage zone": inside $\mathcal{Q}_T$, several edges with positive length under the true model are entirely contracted by AML. More precisely, as shown in Fig. 1, our main result (see Proposition 6) implies that when the true tree has $n-1$ long leaf edges and all other edges are short, AML is guaranteed to return a star rooted at a leaf, as the sequence length goes to infinity. Although our results apply only to a very specific region in the parameter space, this "shrinkage" effect—or more generally, the systematic underestimation of branch lengths—may in fact be common in AML. (See Section 1.3 for some "hints.") We leave for future work the study of the extent of this phenomenon.

We note that our result does not imply the stronger statement that AML is "positively misleading," that is, AML does not return an *incorrect* reconstruction in this case since we can think of the rooted star as the *correct* tree T, where several edges are set to $p_e = 0$. Note, however, that the solution is highly degenerate since the star can be obtained in this way from any tree. In other words, in our "shrinkage zone," AML provides no information about the internal structure of the tree even with infinitely long sequences. Whether or not AML can actually be positively misleading is an interesting question for future work.

1.3 Proof Sketch

The basic idea behind the proof is to consider a tree where all branches are long. In order to obtain some intuition about the construction, it is useful to consider first the extreme case where all branches are of *infinite* length. Clearly, then, the distribution of characters generated at the leaves is simply the uniform distribution on $\{0,1\}^n$.

As we show in the next section (see Definition 6), it follows from Definition 3 that in this case, the AML problem boils down to finding a tree and a distribution of states on the vertices with uniform marginal on the leaves such that the expression

$$\sum_e H(Z_e)$$

is minimized, where Z_e denotes the random variable taking value 1 if there is a substitution on edge e and taking value 0 otherwise. One should think of this formulation simply as a large-k limit of Definition 3.

Note that for the *true* generating process, we have $P[Z_e = 1] = 1/2$ for all $e \in E$ and that the Z_es are independent. Note furthermore that we then have

$$\sum_e H(Z_e) = 2n - 3 = \#\text{ of edges}.$$

It is natural to ask if this is the most "efficient" way to generate the uniform distribution in terms of the quantity $\sum_e H(Z_e)$. A moment's reflection reveals that there are in fact better ways. In particular, it suffices to let all Z_es to be identically 0 except for $n-1$ of the edges pendant at the leaves. In other words, all internal states are taken to be equal to that of a fixed leaf. It is easy to see that in this case, the generated distribution is uniform, and yet

$$\sum_e H(Z_e) = n - 1 < 2n - 3.$$

See Fig. 2 for an illustration of the four-leaf case. Using properties of the entropy function, it is further possible to establish that any assignment of the Z_es consistent with the uniform distribution on the leaves must satisfy

$$\sum_e H(Z_e) \geq n - 1.$$

Note that this shows a "shrinkage" for AML. While the generating tree has infinitely long branches and, therefore, $H(Z_e) = 1$ for all branches, the AML tree has infinitely long branches, and $H(Z_e) = 1$ *only* for $n-1$ of the branches while all other branches have $H(Z_e) = 0$, that is, the branches are contracted to length 0.

The main result of the paper establishes the same kind of phenomenon when the branches of the generating tree are long but of finite length. It shows that in that case, the AML tree has $n-1$ long branches, and all other branches are entirely contracted.

It may look surprising that the branches shrink to length 0 rather than just a *small—but positive—length.* As we show in the next section (see Proposition 1 and the discussion after its proof), *optimal* branch lengths in AML can only take *finitely* many values, even as k tends to $+\infty$: indeed, by a convexity argument, we prove that optimal

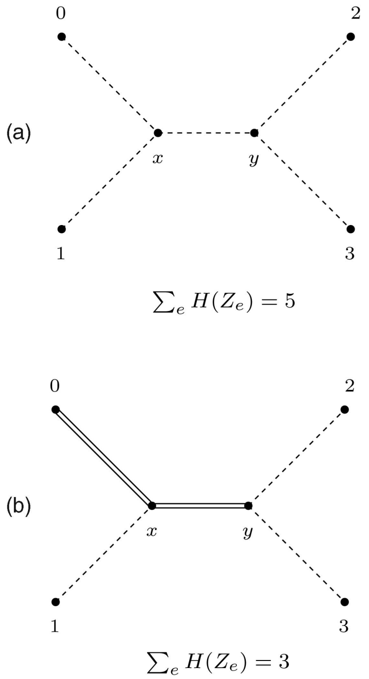


Fig. 2. The dashed edges are infinitely long. The double edges have length 0. (a) The original generating tree. (b) This tree has the same character distribution at the leaves, that is, uniform. Yet, the AML score of the bottom tree is strictly smaller.

branch lengths *must* be consistent with a *fixed* assignment of ancestral states for each possible character (that is, an assignment of internal sequences such that if the same character appears more than once in the data, it must be assigned the same internal states). This implies that for a large k, the only values allowed for the optimal branch lengths are quantities that can be expressed as sums of character probabilities under the generating model—such quantities are independent of k (see the discussion after Proposition 1 for more details). As a consequence, it is *not* possible for the optimal solution to assign *arbitrarily* small edge lengths. The smallest positive value allowed is in fact the smallest positive character probability—which can be much larger than $1/k$ for a large k.

A more detailed analysis allows us to show which of the branches should be contracted and to conclude that the AML tree has a star topology (under the assumptions of our main theorem). See also Fig. 1.

In fact, a weaker result follows immediately from the argument above: branch length estimation in AML cannot be consistent. To see this, consider the four-leaf example in Fig. 3, where all leaf edges are very long, and the internal edge is very short. For concreteness, we say that the internal edge has a substitution probability of 1 percent. Then, since each character appears with probability roughly 1/16, a fixed ancestral assignment to each character can only generate probabilities of substitution that are roughly multiples of 1/16. Indeed, there are $2^4 = 16$ possible character states $0000, 0001, 0010, \ldots, 1111$, each appearing with roughly uniform probability. In particular, following our previous analysis, a small strictly positive value such as 1 percent cannot possibly be achieved by AML in this case (assuming that the topology is correctly reconstructed).

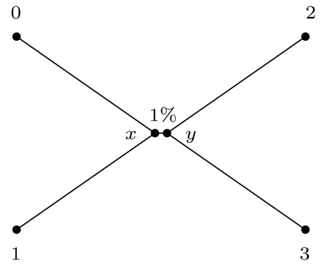


Fig. 3. In this example, the leaf edges are sufficiently long that the character distribution is roughly uniform.

Note finally that ML is not affected by this "shrinkage" phenomenon as it associates the same score to all models with the same distribution of characters at the leaves, and it is not "constrained" to (optimal) substitution probabilities corresponding to discrete ancestral assignments.

1.4 Organization

We begin with some preliminary remarks in Section 2. The proof of Theorem 1 can be found in Section 3.

2 Preliminaries

2.1 Solution Properties

2.1.1 Fixed Extension

Let $T \in \mathcal{T}_n$. For an assignment of sequences $\boldsymbol{\mu} \in \mathcal{B}_n^{(k)}$ and $1 \leq j \leq k$, we call $\chi : [n] \rightarrow \{0,1\}$ with $\chi_u = (\mu_u)_j$ for all $u \in [n]$ the jth *character* in $\boldsymbol{\mu}$. We write $\chi \in \boldsymbol{\mu}$ if there is a j such that χ is the jth character in $\boldsymbol{\mu}$. We also denote by $\chi^{\#}$ the number of characters in $\boldsymbol{\mu}$ equal to χ. An extension of a character χ is a mapping $\bar{\chi} : V \rightarrow \{0,1\}$ such that $\bar{\chi}_v = \chi_v$ for all $v \in [n]$. We denote by $\mathcal{V}(\chi)$ the set of all extensions of χ on T. Let $f : \{0,1\}^{[n]} \rightarrow \{0,1\}^{V-[n]}$. The mapping f then defines an extension $\bar{\chi}_f$ for all characters χ simultaneously by setting $(\bar{\chi}_f)_v = \chi_v$ for all $v \in [n]$ and $(\bar{\chi}_f)_v = f(\chi)_v$ for all $v \in V - [n]$. In other words, for each character χ, an extension f assigns a unique state in {0, 1} to each internal node in V. We show next that AML is in fact equivalent to finding such an f—which can significantly reduce the size of the problem for a large k as the number of possible characters is finite. For a set of n binary sequences, $\boldsymbol{\mu} \in \mathcal{B}_n^{(k)}$ and a tree $T = (V, E) \in \mathcal{T}_n$, we denote by $\bar{\boldsymbol{\mu}}_f$ the extension of $\boldsymbol{\mu}$ to V by applying f as above to every character in $\boldsymbol{\mu}$.

Definition 5 (AML version 3). *Given a set of n binary sequences $\boldsymbol{\mu} \in \mathcal{B}_n^{(k)}$, find a tree $T \in \mathcal{T}_n$ and a mapping $f : \{0,1\}^{[n]} \rightarrow \{0,1\}^{V-[n]}$ such that the quantity*

$$\mathcal{H}(T \mid \bar{\boldsymbol{\mu}}_f) = \sum_{e \in E} H\left(\frac{d_e}{k}\right)$$

is minimized, where

$$d_{u,v} = \left\|(\bar{\mu}_f)_u - (\bar{\mu}_f)_v\right\|_1.$$

Note that in AML version 2 (see Definition 3), if the same character appears more than once, it can be mapped to *different* internal assignments. This is explicitly *forbidden* in AML version 3, as justified by the next proposition.

Proposition 1 (AML, fixed extension). *There is always a solution of AML version 1 (and 2) of the form $\boldsymbol{\lambda} = \bar{\boldsymbol{\mu}}_f$ for some $f : \{0,1\}^{[n]} \to \{0,1\}^{V-[n]}$ with corresponding assignment $\mathbf{p} : E \to [0,1]$ of edge probabilities, where*

$$p_e = \frac{d_e}{k},$$

for all e with

$$d_{u,v} = \left\|(\bar{\mu}_f)_u - (\bar{\mu}_f)_v\right\|_1.$$

Moreover, all optimal assignments of edge probabilities under AML version 1 must be of the form above, for some fixed extension.

Proof. As before, we denote the data set by $\boldsymbol{\mu}$. Let T^*, $\mathbf{p}^*$, and $\boldsymbol{\lambda}^*$ be any optimal solution under AML version 1. Note that

$$\begin{aligned}\mathcal{L}(T, \mathbf{p} \mid \boldsymbol{\lambda}) &= -\log_2\left(\prod_{e\in E} p_e^{d_e}(1-p_e)^{k-d_e}\right)\\ &= -\sum_{e\in E}\log_2\left(p_e^{d_e}(1-p_e)^{k-d_e}\right)\\ &= -k\sum_{e\in E}\log_2(1-p_e)\\ &\quad -\sum_{j=1}^{k}\sum_{(u,v)\in E}\mathbb{1}\left\{(\lambda_u)_j \neq (\lambda_v)_j\right\}\log_2\frac{p_e}{1-p_e},\end{aligned}$$

where $\mathbb{1}\{\mathcal{E}\} = 1$ if $\mathcal{E}$ is satisfied and is 0 otherwise. We make two observations:

1. From the third equality, we deduce that for fixed T and $\mathbf{p}$, the quantity $\mathcal{L}$ "decomposes linearly" in j. Hence, it is always possible to modify $\boldsymbol{\lambda}^*$ so as to take the *same* extension for each character appearing in the data $\boldsymbol{\mu}$, without affecting optimality. Let f^* be such a fixed extension that is optimal under T^*, $\mathbf{p}^*$. That is, we let $\boldsymbol{\lambda}^* = \bar{\boldsymbol{\mu}}_{f^*}$.
2. It was noted in [1] that the expression $p_e^{d_e}(1-p_e)^{k-d_e}$ as a function of p_e is uniquely maximized at $p_e = d_e/k$. Therefore, given T^* and $\bar{\boldsymbol{\mu}}_{f^*}$, we deduce from the second equality that the unique edge probability maximizer $\mathbf{p}^*$ *must* satisfy

$$p_e^* = \frac{d_e^*}{k}, \tag{3}$$

for all e with

$$d_{u,v}^* = \left\|(\bar{\mu}_{f^*})_u - (\bar{\mu}_{f^*})_v\right\|_1. \tag{4}$$

In other words, if $\mathbf{p}^*$ is not of this form, we get a contradiction by strictly improving the solution using the above solution form.

The solution sought in the statement of the proposition is given by T^*, $\mathbf{p}^*$, and $\bar{\boldsymbol{\mu}}_{f^*}$. The second claim immediately follows from Point 2 above. □

The second claim of Proposition 1 says that *any optimal* edge probability assignment $\mathbf{p}^*$ *must* correspond to a fixed extension f^*. As we discussed informally in Section 1.3, this has an important consequence. Assume that $\mathbf{p}^*$ is defined as in (3) and (4). Note that for $e = (u, v)$, we have

$$\begin{aligned}\frac{d_e^*}{k} &= k^{-1}\left\|(\bar{\mu}_{f^*})_u - (\bar{\mu}_{f^*})_v\right\|_1\\ &= \sum_{\chi\in\boldsymbol{\mu}}\left(\frac{\chi^{\#}}{k}\right)\mathbb{1}\{\bar{\chi}_u^* \neq \bar{\chi}_v^*\},\end{aligned}$$

where $\bar{\chi}^*$ is the extension of χ under $\bar{\boldsymbol{\mu}}_{f^*}$. By the Law of Large Numbers, $\frac{\chi^{\#}}{k}$ converges to the probability of observing the character χ under the generating model, a quantity *independent of* k. Hence, for a large k (and fixed n), $\frac{d_e^*}{k}$ can be written as a simple combination of character probabilities under the true model—restricting significantly its possible values. In particular, the smallest nonzero value allowed for the optimal p_e^* is the smallest nonzero character probability under the true model, which can be much larger than $1/k$ for a large k.

2.1.2 Limit Problem

Let $T = (V, E) \in \mathcal{T}_n$. Assume as in Theorem 1 that we are given a data set $\mathbb{X} = \{X_1, X_2, \ldots\}$ with $X_1, X_2, \ldots$ i.i.d. $\mathrm{CFN}(T, \mathbf{p})$. Fix $f : \{0,1\}^{[n]} \to \{0,1\}^{V-[n]}$. Let $X \sim \mathrm{CFN}(T, \mathbf{p})$ and denote by $Y = \bar{X}_f$ the extension of X under f. Also, let $\bar{\boldsymbol{\mu}}^{(k)}_{\mathbb{X},f}$ be the extension of $\boldsymbol{\mu}^{(k)}_{\mathbb{X}}$ under f. By the Law of Large Numbers, as $k \to +\infty$, the quantity $\mathcal{H}(T|\bar{\boldsymbol{\mu}}^{(k)}_{\mathbb{X},f})$ converges almost surely to

$$\mathbb{H}_{X,T}(f) = \sum_{e\in E} H(Z_e),$$

where for $e = (u, v)$, Z_e is the indicator that $Y_u \neq Y_v$, and $H(Z_e)$ is the entropy of Z_e, that is

$$H(Z_e) = H(\mathbb{P}[Y_u \neq Y_v]).$$

Note that by Proposition 1, even as $k \to +\infty$, there are only a constant number of mappings f to consider. We say that f is $\mathbb{H}_{X,T}$-optimal if f minimizes $\mathbb{H}_{X,T}(f)$ over all $f : \{0,1\}^{[n]} \to \{0,1\}^{V-[n]}$. The minimum need not be unique.

Definition 6 (expected AML). *Given a random variable X taking values in $\{0,1\}^{[n]}$, find a tree $T = (V, E) \in \mathcal{T}_n$ and a fixed extension $f : \{0,1\}^{[n]} \to \{0,1\}^{V-[n]}$ such that the quantity*

$$\mathbb{H}_{X,T}(f) = \sum_{e\in E} H(Z_e)$$

is minimized, where $(Z_e)_{e\in E}$ is as above with $Y = \bar{X}_f$.

By the previous remarks and (2), to prove Theorem 1, it suffices to show the following.

Theorem 2 (optimal assignment). *Let $T' = (V', E') \in \mathcal{T}_n$ and let $X \sim \mathrm{CFN}(T', \mathbf{p})$. Then, there is a constant $\beta > 0$ and a set $\mathcal{Q}_{T'} = \prod_{e\in E'} I_e$ with $|I_e| > \beta$ such that for all $\mathbf{p} \in \mathcal{Q}_{T'}$ and $T = (V, E) \in \mathcal{T}_n$, the unique $\mathbb{H}_{X,T}$-optimal $f : \{0,1\}^{[n]} \to \{0,1\}^{V-[n]}$ assigns to all internal nodes of V the value at leaf 0 under all characters, that is*

$$f(x) = (x_0, \ldots, x_0),$$

for all $x \in \{0,1\}^{[n]}$.

2.2 Minimal Isolating Sets

2.2.1 Definition

In preparation for our proof of Theorem 2, we will need the following notion, which is studied in [13].

Definition 7 (isolating set). *Let $T = (V, E)$ be a tree. A subset S of E is called an* isolating set *for T if for any two leaves u and v, there exists an edge $e \in S$ on the path connecting u and v.*

The following result is proved in [13].

Proposition 2 (minimal isolating set). *The size of a minimal isolating set on an n-leaf tree is $n - 1$.*

We will also need the following proposition.

Proposition 3 (one leaf per component). *Let T be a tree on n leaves and let S be a minimal isolating set on T. Consider the forest F obtained from T by removing all edges in S. Then, each component of F contains exactly one leaf of T.*

Proof. If a component of F contains two leaves, then these cannot be isolated under S, a contradiction. On the other hand, if a component T' of F does not contain a leaf, then every edge adjacent to T' in T is in fact in S. But then, one can remove one of these edges without losing the isolating property of S, contradicting the minimality of S. □

2.2.2 Minimally Isolating f

Let $T = (V, E) \in \mathcal{T}_n$ and $f : \{0,1\}^{[n]} \to \{0,1\}^{V-[n]}$. We denote by $S_f \subseteq E$ the set of edges $e = (u, v)$ such that there is $x \in \{0,1\}^{[n]}$ with $f(x)_u \neq f(x)_v$.

Definition 8 (minimally isolating f). *We say that f is minimally isolating for T if S_f is a minimal isolating set of T.*

2.3 Random Cluster Parameterization

We will sometimes require a different ("random cluster") parameterization of the CFN model. Let $T \in \mathcal{T}_n$ and $\mathbf{p} \in [0,1]^E$. (Note that we allow p_e in $[0,1]$.) We let

$$\theta_e = 1 - 2p_e,$$

for all $e \in E$. The main property we will use is the following well-known identity. For two leaves u and v in T, let $\mathrm{Path}_T(u, v)$ be the set of edges on the path between u and v.

Proposition 4 (path probability). *Let $T = (V, E) \in \mathcal{T}_n$ and $\mathbf{p} \in [0,1]^E$. Assume that $X \sim \mathrm{CFN}(T, \mathbf{p})$. Let u and v be two leaves of T. Then, it is well known (see, e.g., [7]) and easily proved by induction that*

$$\mathbb{P}[X_u \neq X_v] = \frac{1}{2}\left(1 - \prod_{e \in \mathrm{Path}_T(u,v)} \theta_e\right).$$

3 Proof

In this section, we prove Theorem 2 from which Theorem 1 follows. The proof has two components:

1. [Reduction to Minimal Isolating Sets] We first show that for any random variable $X \in \{0,1\}^{[n]}$ close enough to uniform and any tree $T \in \mathcal{T}_n$, the $\mathbb{H}_{X,T}$-optimal f's are minimally isolating for T.
2. [Rooted Star Is Optimal] Second, we show that if X above is $\mathrm{CFN}(T', \mathbf{p})$ for some $T' \in \mathcal{T}_n$ with $p_e \approx 1/2$ if e is adjacent to $\{1, \dots, n-1\}$ and $p_e \approx 0$ otherwise, then for all $T \in \mathcal{T}_n$, the unique $\mathbb{H}_{X,T}$-optimal f assigns the value at 0 to all internal nodes.

Throughout, $n \geq 1$ is fixed.

3.1 Reduction to Minimal Isolating Sets

We prove the following.

Proposition 5 (reduction to minimal isolating sets). *There exists $\varepsilon > 0$ (depending on n) such that the following hold. Let X be any random variable taking values in $\{0,1\}^{[n]}$ with $H(X) \geq n - \varepsilon$ and let T be any tree in $\mathcal{T}_n$. If f is $\mathbb{H}_{X,T}$-optimal, then f is minimally isolating for T.*

Proof. We make a series of claims. □

Claim 1 (reduction to uniform). *For all $\delta > 0$, there exists $\varepsilon = \varepsilon(\delta) > 0$ such that if X is a $\{0,1\}^{[n]}$-random variable with*

$$H(X) \geq n - \varepsilon$$

and $f : \{0,1\}^{[n]} \to \{0,1\}^{V-[n]}$, then

$$\left|\mathbb{H}_{X,T}(f) - \mathbb{H}_{U,T}(f)\right| \leq \delta, \tag{5}$$

where U is the uniform distribution on $\{0,1\}^{[n]}$. Therefore, it suffices to prove Proposition 5 for those f that are $\mathbb{H}_{U,T}$-optimal.

Proof. The entropy of $\{0,1\}^{[n]}$-random variables is maximized uniquely at $H(U) = n$. The first part of the result follows by continuity of $H(X)$ and $\mathbb{H}_{X,T}(f)$ in the distribution of X.

For the second part, take a small enough $\delta > 0$ such that for all f and f', we have

$$\mathbb{H}_{U,T}(f) > \mathbb{H}_{U,T}(f') \Longrightarrow \mathbb{H}_{U,T}(f) > \mathbb{H}_{U,T}(f') + 2\delta. \tag{6}$$

(Recall that there are only finitely many f's for fixed n.) Take $\varepsilon > 0$ such that the first part holds. Then, it follows that if f is $\mathbb{H}_{X,T}$-optimal, then it must be $\mathbb{H}_{U,T}$-optimal. We argue by contradiction. Assume that there are f and f' such that $\mathbb{H}_{X,T}(f) \leq \mathbb{H}_{X,T}(f')$, but $\mathbb{H}_{U,T}(f) > \mathbb{H}_{U,T}(f')$. By (6), we have

$$\mathbb{H}_{U,T}(f) > \mathbb{H}_{U,T}(f') + 2\delta, \tag{7}$$

which implies $\mathbb{H}_{X,T}(f) > \mathbb{H}_{X,T}(f')$ by (5), a contradiction. □

Claim 2 (minimizer). *If f is $\mathbb{H}_{U,T}$-optimal, then $\mathbb{H}_{U,T}(f) = n - 1$. Moreover, denoting $Y = \bar{U}_f$, we have that $\{Y_0, (Z_e)_{e \in E}\}$ are mutually independent.*

Proof. *Upper bound.* We first show that there is f such that $\mathbb{H}_{U,T}(f) \leq n - 1$. Let S be a minimal isolating set for T. Define f by letting $f(x)_u = f(x)_v$ for all edges (u, v) not in S. By Proposition 3, this uniquely defines f. Letting $Y = \bar{U}_f$, it is immediate to check that

$$\mathbb{H}_{U,T}(Y) = \sum_{e \in E} H(Z_e) = \sum_{e \in S} H(Z_e) \leq n - 1,$$

by Proposition 2.

Lower bound. For any $f : \{0,1\}^{[n]} \to \{0,1\}^{V-[n]}$ with $Y = \bar{U}_f$, we have

$$\begin{aligned} n = H(U) &= H\Big(\Big\{(Y_v)_{v\in[n]}\Big\}\Big) = H\big(\{Y_0, (Z_e)_{e\in E}\}\big) \\ &\le H(Y_0) + \sum_{e\in E} H(Z_e) \le 1 + \sum_{e\in E} H(Z_e), \end{aligned}$$

where we have used that $\{(Y_v)_{v\in[n]}\}$ and $\{Y_0, (Z_e)_{e\in E}\}$ are deterministic functions of each other. Furthermore, the first inequality holds to equality if and only if $\{Y_0, (Z_e)_{e\in E}\}$ are mutually independent. □

We are ready to conclude the proof of Proposition 5. Let f be $\mathbb{H}_{U,T}$-optimal with $Y = \bar{U}_f$. Let u and v be any two leaves of T. We have by the previous claim that $(Z_e)_{e\in\mathrm{Path}_T(u,v)}$ are mutually independent. Since Y_u and Y_v are independent and uniform in {0, 1}, it must be that there is an edge $e \in \mathrm{Path}_T(u,v)$ with $H(Z_e) = 1$. Indeed, define $p_e = \mathbb{P}[Z_e = 1]$ and $\theta_e = 1 - 2p_e$. Then, by Proposition 4, we have

$$0 = 1 - 2\mathbb{P}[Y_u \neq Y_v] = \prod_{e\in\mathrm{Path}_T(u,v)} \theta_e,$$

which implies that at least one $\theta_e = 0$. Let S' be the set of all edges where $H(Z_e) = 1$. Then, we have shown that S' is an isolating set. Note furthermore that if $e \in S_f$, then $H(Z_e) \ge H(2^{-n}) > 0$. From f's optimality, we obtain

$$n - 1 = \mathbb{H}_{U,T}(f) \ge |S'| + |S_f \setminus S'| H(2^{-n}).$$

Therefore, we must have $S_f = S'$ and $|S'| = n - 1$, which implies that S_f is a minimal isolating set as needed. □

3.2 The Rooted Star Is Optimal

Let $T = (V, E) \in \mathcal{T}_n$ and S be a minimal isolating set of T. Let T^0 be the tree obtained from T by contracting all edges not in S. By Proposition 3, T^0 is an n-node tree where each node (leaf or internal) is (uniquely) labeled by a leaf of T. Let $\mathcal{T}^0_n$ be all such trees on n nodes. By Proposition 5, the AML phylogeny estimator is among $\mathcal{T}^0_n$. Note that for $T \in \mathcal{T}^0_n$, the only possible extension is the identity $f = \mathrm{Id}$ since there are no unlabeled internal vertices.

Proposition 6 (rooted star is optimal). *Let $T = (V, E) \in \mathcal{T}_n$. Let W be the set of leaf edges of T, except the edge pendant at leaf 0. Then, for sufficiently small $\varepsilon, \delta > 0$, the following holds. Assume that $X \sim \mathrm{CFN}(T, \mathbf{p})$ with corresponding random cluster parameterization satisfying $0 < \theta_e \le \varepsilon$ for all $e \in W$ and $1 > \theta_e > 1 - \delta$ for all $e \notin W$. Then, among all trees $T' \in \mathcal{T}^0_n$, the star rooted at 0 uniquely minimizes $\mathbb{H}_{X,T'}(\mathrm{Id})$ for all sufficiently small δ.*

Proof. We assume that δ and ε are small enough so that they satisfy

$$(n-1)(1-\delta)^{2n-4} > n - 2$$

and

$$\varepsilon^2 < (n-1)(1-\delta)^{2n-4} - (n-2). \qquad (8)$$

Let $T' = (V', E') \in \mathcal{T}^0_n$ and $f = \mathrm{Id}$ with corresponding variables $(Y_0, \{Z_e\}_{e\in E'})$, where $Y_0 = X_0$, and $Z_{u,v} = \mathbb{1}\{X_u \neq X_v\}$. Let $e = (u, v)$ be an edge in T'. In particular, note that u and v are leaves of T. Let $p_{u,v}$ be the probability that u and v disagree and let $\theta_{u,v} = 1 - 2p_{u,v}$. We will use the following Taylor expansion of the entropy around $1/2$:

$$H\left(\frac{1-\tau}{2}\right) = 1 - \left(\frac{\log_2 e}{2}\right)\tau^2 + O(\tau^4).$$

Note further that

$$H(Z_e) = H(p_{u,v}) = H\left(\frac{1-\theta_{u,v}}{2}\right).$$

As ε approaches 0, $p_{u,v}$ goes to 1/2. Therefore, by Proposition 4, up to smaller order terms, we want to find $T' = (V', E')$ in $\mathcal{T}^0_n$ that maximizes

$$\Theta(T') := \sum_{e'=(u,v)\in E'} \prod_{e\in\mathrm{Path}_T(u,v)} \theta_e^2.$$

If T' has an edge e' between two leaves neither of which is 0, then e' makes a contribution of at most ε^4 to $\Theta(T')$ since $\mathrm{Path}_T(u,v)$ crosses two edges in W. Therefore, by (8), we have

$$\begin{aligned} \Theta(T') &\le (n-2)\varepsilon^2 + \varepsilon^4 \\ &< (n-1)(1-\delta)^{2n-4}\varepsilon^2, \end{aligned}$$

where we have used that T' has exactly $n-1$ edges, and each edge $e'' \in E'$ makes a contribution of at most ε^2 since $\mathrm{Path}_T(u,v)$ contains at least one edge in W. On the other hand, the star rooted at 0, which we denote by T^*, is the only tree in $\mathcal{T}^0_n$ that does not include an edge between two leaves neither of which is 0. In that case, we get

$$\Theta(T^*) \ge (n-1)(1-\delta)^{2(n-2)}\varepsilon^2,$$

where we have used that any path between 0 and another leaf in T contains at most $n-2$ edges not in W (since $|E| \le 2n-3$ and $|W| = n-1$) and exactly one edge in W. Taking a small enough ε gives the result. □

4 Concluding Remarks

It would be interesting to extend our results beyond the two-state case. We note in particular that for the symmetric r-state model, with $r > 2$, the equivalent formulation of the AML problem given in Definition 3 does not apply. Indeed, it is easy to check that instead, one needs to minimize

$$\mathcal{H}'(T \mid \boldsymbol{\lambda}) = \sum_{e\in E} H\left(\frac{d_e}{k}\right) + \log_2(r-1)\sum_{e\in E}\frac{d_e}{k}.$$

The second term on the right-hand side—a parsimony "correction"—may lead to a different behavior when $r > 2$.

We thank Peter Ralph for sharing his recent independent results [16] regarding the structure of the optimal solution in the two-state case (similar to [2]), as well as a number of simulations on four-taxon trees.

ACKNOWLEDGMENTS

Elchanan Mossel was supported by a Sloan Fellowship in Mathematics, US National Science Foundation (NSF) Grant DMS 0528488, NSF CAREER Award DMS 0548249, and ONR Grant N0014-07-1-05-06. The authors thank the anonymous referees for their helpful suggestions and comments.

REFERENCES

[1] L. Addario-Berry, B. Chor, M. Hallett, J. Lagergren, A. Panconesi, and T. Wareham, "Ancestral Maximum Likelihood of Evolutionary Trees Is Hard," *J. Bioinformatics and Computational Biology,* vol. 2, no. 2, pp. 257-271, 2004.

[2] N. Alon, B. Chor, F. Pardi, and A. Rapoport, "Approximate Maximum Parsimony and Ancestral Maximum Likelihood," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* to appear.

[3] D. Barry and J. Hartigan, "Statistical Analysis of Hominoid Molecular Evolution," *Statistical Science,* vol. 2, pp. 191-207, 1987.

[4] D. Barry and J. Hartigan, "Rejoinder [On Statistical Analysis of Hominoid Molecular Evolution]," *Statistical Science,* vol. 2, pp. 209-210, 1987.

[5] J.A. Cavender, "Taxonomy with Confidence," *Math. Biosciences,* vol. 40, nos. 3/4, 1978.

[6] B. Chor and T. Tuller, "Finding the Maximum Likelihood Tree Is Hard," *Proc. Ninth Ann. Int'l Symp. Research in Computational Biology (RECOMB),* 2005.

[7] P.L. Erdös, M.A. Steel, L.A. Székely, and T. Warnow, "A Few Logs Suffice to Build (Almost) All Trees (Part 1)," *Random Structures and Algorithms,* vol. 14, no. 2, pp. 153-184, 1999.

[8] J.S. Farris, "A Probability Model for Inferring Evolutionary Trees," *Systematic Zoology,* vol. 22, no. 4, pp. 250-256, 1973.

[9] J. Felsenstein, "Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading," *Systematic Biology,* vol. 27, pp. 401-410, 1978.

[10] J. Felsenstein, *Inferring Phylogenies.* Sinauer, 2004.

[11] N. Goldman, "Maximum Likelihood Inference of Phylogenetic Trees, with Special Reference to a Poisson Process of DNA Substitution and to Parsimony Analysis," *Systematic Zoology,* vol. 39, pp. 345-361, 1990.

[12] P.A. Goloboff, "Parsimony, Likelihood, and Simplicity," *Cladistics,* vol. 19, pp. 91-103, 2003.

[13] V. Moulton and M. Steel, "Peeling Phylogenetic 'Oranges'," *Advances in Applied Math.,* vol. 33, no. 4, pp. 710-727, 2004.

[14] J. Neyman, "Molecular Studies of Evolution: A Source of Novel Statistical Problems," *Statistical Decision Theory and Related Topics,* S.S. Gupta and J. Yackel, eds., pp. 1-27, Academic Press, 1971.

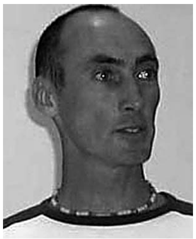
[15] T. Pupko, I. Pe'er, R. Shamir, and D. Graur, "A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences," *Molecular Biology and Evolution,* vol. 17, no. 6, pp. 890-896, 2000.

[16] P. Ralph, in preparation, 2008.

[17] S. Roch, "A Short Proof That Phylogenetic Tree Reconstruction by Maximum Likelihood Is Hard," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 3, no. 1, pp. 92-94, Jan.-Mar. 2006.

Elchanan Mossel received the PhD degree in mathematics from the Hebrew University. After completing his PhD, he was a postdoctoral researcher in the Theory Group, Microsoft Research, and a Miller fellow in statistics and computer science at the University of California, Berkeley. He is currently on the faculty of the Department of Statistics and Department of Computer Science, University of California, Berkeley, and a member of the Faculty of Mathematics and Computer Science, Weizmann Institute. He was awarded a Sloan Fellowship in Mathematics and a US National Science Foundation CAREER Award.

Sebastien Roch received the PhD degree in statistics from the University of California, Berkeley, under the guidance of Professor Elchanan Mossel. He is currently a postdoctoral researcher at Microsoft Research, Cambridge, Massachusetts.



Mike Steel studied mathematics at the University of Canterbury, Christchurch, New Zealand, and received the PhD degree in mathematics from Massey University, New Zealand, in 1989. From 1990 to 1993, he held various postdoctoral positions in Germany and New Zealand and was appointed to a tenured position at the University of Canterbury in 1994. He is currently a professor in the Department of Mathematics and Statistics and the director of the Biomathematics Research Centre at the University of Canterbury and is a principal investigator in the Allan Wilson Centre for Molecular Ecology and Evolution.