Finding the Maximum Likelihood Tree is Hard *

(Submitted to JACM, July 6, 2005)

Benny Chor † Tamir Tuller ‡

School of Computer Science
Tel-Aviv University

Abstract

Maximum likelihood (ML) is an increasingly popular optimality criterion for selecting evolutionary trees (Felsenstein, 1981). Finding optimal ML trees appears to be a very hard computational task, but for tractable cases, ML is the method of choice. In particular, algorithms and heuristics for ML take longer to run than algorithms and heuristics for the second major character based criterion, maximum parsimony (MP). However, while MP has been known to be NP-complete for over 20 years (Graham and Foulds, 1982; Day, Johnson, and Sankoff, 1986), such a hardness result for ML has so far eluded researchers in the field.

An important work by Tuffley and Steel (1997) proves quantitative relations between the parsimony values of given sequences and the corresponding log likelihood values. However, a direct application of their work would only give an *exponential time* reduction from MP to ML. Another step in this direction has recently been made by Addario-Berry *et al.* (2004), who proved that *ancestral maximum likelihood* (AML) is NP-complete. AML "lies in between" the two problems, having some properties of MP and some properties of ML. Still, the AML proof is not directly applicable to the ML problem.

We resolve the question, showing that "regular" ML on phylogenetic trees is indeed intractable. Our reduction follows the vertex cover reductions for MP (Day *et al.*) and AML (Addario-Berry *et al.*), but its starting point is an approximation version of vertex cover, known as GAP VC. The crux of our work is not the reduction, but its correctness proof. The proof goes through a series of tree modifications, while controlling the likelihood losses at each step, using the bounds of Tuffley and Steel. The proof can be viewed as correlating the value of any ML solution to an arbitrarily close approximation to vertex cover.

Key words: Maximum likelihood, tree reconstruction, maximum parsimony, intractability, approximate vertex cover.

1 Background

Molecular data, and even complete genomes, are being sequenced at an increasing pace. This newly accumulated information should make it possible to resolve long standing questions in evolution, such as reconstructing the phylogenetic tree of placental mammals and estimating the times of species divergence. The analysis of this data flood requires sophisticated mathematical tools and algorithmic techniques. Two character-based methods are widely used in practice: MP

*Research supported by ISF grant 418/00. An extended abstract of this work was submitted to RECOMB2005 on October 29, 2004, and published in the conference proceedings on May 14, 2005.

†benny@cs.tau.ac.il

‡corresponding author, tamirtul@post.tau.ac.il

(*maximum parsimony*, Fitch, 1971 [11]) and ML (*maximum likelihood*, Felsenstein, 1981 [8]). It is known that ML is *consistent*, namely with high probability, for long enough input sequences, the correct tree is the tree maximizing the likelihood [10, ch. 16]. Consistency does not hold for MP, and in fact for certain families of trees (the so called *Felsenstein zone* [9]) MP will reconstruct the *wrong* trees, even for arbitrarily long input sequences. The two methods are known to be computationally intensive, and exact algorithms are limited to just about $n = 20$ sequences. This forces practitioners to resort to heuristics. For both exact algorithms and heuristics, ML seems a *harder* problem than MP.

In the absence of concrete lower bound techniques, the major tool for demonstrating computational intractability remains NP hardness proofs. Both MP and ML have well-defined objective functions, and the related decision problems (or at least discretized versions of them) are in the complexity class NP. It has been known for over 20 years that MP is NP-complete [13, 5, 12, 4, 6, 19], (see also [23] and references). The proof of [5] employs an elegant reduction from vertex cover (VC). However, no such result has been found for ML to date. This is particularly frustrating in light of the intuition among practitioners that ML is harder than MP.

Tuffley and Steel have investigated the quantitative relations between MP and ML [22]. In particular, they showed that if the n sequences are padded with sufficiently many zeroes, the ML and MP trees coincide. Since parsimony is invariant under padding by zeroes, this approach could in principle lead to a reduction from MP to ML. Unfortunately, the upper bound provided in [22] on the padding length is *exponential* in n. A step in a different direction was taken by Addario-Berry *et al.* [1]. They studied the complexity of AML (Ancestral Maximum Likelihood) [17, 24]. This variant of ML is "between" MP and ML in that it is a likelihood method (like ML) but it reconstructs sequences for internal vertices (like MP). They showed that AML is NP-complete, using a reduction from (exact) Vertex Cover.

Our NP hardness proof of ML uses ingredients from both [22] and [1], as well as new insights on the behavior of the likelihood function on trees. The reduction itself is essentially identical to that given for MP by Day, Johnson, and Sankoff [5], and also used in the AML paper [1]. However, our starting point is not *exact* VC but the *gap* version of it [2, 16]. The proof of correctness for this reduction relative to ML is different, and substantially more involved. We define a family of *canonical trees*. Every such tree is associated with a unique cover in the original graph. We show that if L is the likelihood of the canonical tree, n is the number of vertices in the original graph, m is the number of edges in the original graph, and c is the size of the associated cover, then as $n \to \infty$,

$$\frac{-\log(L)}{(m+c)\log(n)} \to 1 \ .$$

In particular, this gives an inverse relation between likelihood and cover size: Larger L implies smaller c, and vice versa.

When proving the correctness of the reduction, we want to establish two directions: ($\Rightarrow$) If the original graph has a small cover, then there is a tree with high likelihood, and ($\Leftarrow$) that the existence of a tree with high likelihood implies the existence of a small cover. The first direction is easy, using the canonical tree related to the small vertex cover. It is the other direction that is hard, because there is no obvious relation between the log likelihood of a *non-canonical* tree and the size of any cover. What we do, starting from any ML tree, is to apply a sequence of modifications that leads it to a *canonical tree*. The whole series of modifications may actually *decrease* the likelihood of the resulting, canonical tree vs. the original, ML one. We use the techniques of [22] to infer likelihood properties from parsimony ones. In particular, we combine [22] and the degree bound of the original graph to show that in every step, the log likelihood decreases by at most $O(\log n)$ bits. Finally, we show that the total number of modifications is not too large – at most $n/\log\log n$. This allows us to show that the overall loss in log likelihood

is at most $O(n \log n / \log \log n)$. We also show that the log likelihood of the final, canonical tree is $\theta(n \log n)$. This implies the ratio of the log likelihood of the last, canonical tree, and the log likelihood of the ML tree, approaches 1 as $n \to \infty$. This proves that log ML is tightly related to an approximate vertex cover, establishing the NP hardness of ML.

2 Proof's Overview

In this section we give a high level description of the hardness proof. The reduction is from the GAP VERTEX COVER problem on graphs whose degree is at most three, a problem proved NP-hard in 1999 by Berman and Karpinski [2, 16].

Given an undirected graph $G = (V, E)$ of max degree 3 with $n = |V|$ nodes and $m = |E| \leq 1.5n$ edges, we construct an ML instance, consisting of $m + 1$ binary strings of length n. The ML problem is to find a tree with the $m+1$ sequences at its leaves, and an assignment of substitution probabilities to the edges of that tree (edges' length), such that the likelihood of generating the given sequences is maximized. The proof relates the approximate max log likelihood value to the size of a vertex cover in G. This approximation is tight enough to enable solving the original gap problem. Our reduction follows the one for maximum parsimony given by Day, Johnson and Sankoff [5] and for ancestral ML, given by Addario-Berry *et al.* [1]. Both reductions were from the (exact) VERTEX COVER problem. In this reduction we generate one string with only 0s, and m "edge strings" that contain exactly two 1s each, and naturally encodes an edge.

Consider all unrooted weighted trees with $m+1$ leaves that have the given sequences at their leaves. We say that such tree is in *canonical form* if the following properties hold (see figure 1):

Definition 2.1

1. *nal node (called the* "root" *for clarity, even though the trees are unrooted) that has the all zero leaf as a son, and the length of the edge going to this leaf is* 0*.*
2. *All leaves are one or two tree edges away from the root.*
3. *If a leaf is two tree edges away from the root, then the subtree that contains that leaf has two or three leaves. In this case, all two or three sequences at the leaves share a "*1*" in the same position.*

Canonical trees uniquely define a vertex cover, where each subtree corresponds to one, two, or three original edges that are covered by one node. (For the subtrees with one leaf, the covering vertex can correspond to either end point, while for size two and three subtrees, the covering vertex is uniquely defined.) Consequently, given a tree in canonical form, we can quantify the size of the corresponding vertex cover of the original graph. The reason we force the root to be connected to the all zero leaf with an edge of weight 0 is that this way the root itself is "effectively forced" to the all zero label (with probability 1). This enables us to express the likelihood of a canonical form tree as a product of the likelihoods of its subtrees. In particular, there is no influence, or dependency, between different subtrees.

The major part of the proof is showing that given any ML tree, T_{ML}, with the given "reduction sequences" at its leaves, there is a series of local modifications on trees with the given sequences at their leaves, such that in each modification the log likelihood of the resulting tree is decreased by at most $O(\log n)$ per step, and the final tree, T_{CA}, is in canonical form. The number of modifications is $o(n)$, which is small enough to establish a tight ratio $1 - o(1)$ between the max log likelihood and the log likelihood of the final, canonical tree. In each step, we transform one tree to another. We identify a small forest, containing between $\log \log n$ and $2 \log \log n$ leaves.

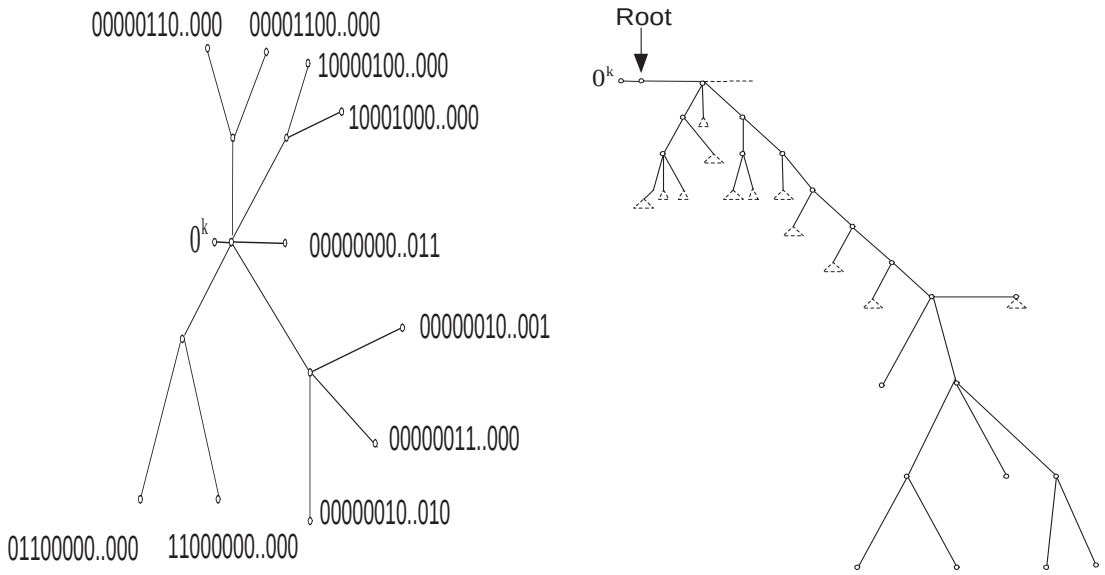

Figure 1: Canonical (left) and non-canonical (right) trees.

Such a forest is a union of disjoint subtrees that are hung off a common internal node (not the "root"). Using the bound on the degree of the original graph, we show that the parsimony score of this forest when its root is labeled by the all zero string can be worse by at most a constant $B < 8^4$ than the score with any other root labeling. Using the results of Tuffley and Steel [22], and the small size of the subtree, it is possible to unroot this forest, rearrange it, and connect it directly to the root in a "canonical way", such that the overall log likelihood of the whole tree decreases by at most $B \log n + o(\log n)$. Over the series of $n/\log\log n$ modifications, the overall decrease is at most $Bn \log n/\log\log n + o(n \log n/\log\log n) = O(n \log n/\log\log n)$. We show that the log likelihood of the final canonical tree, T_{CA}, is $\theta(n \log n)$. This is sufficiently large to show that despite such decrease,

$$\frac{\log L(S|T_{CA})}{\log L(S|T_{ML})} = 1 - o(1) \ .$$

Every tree in canonical form naturally corresponds to a vertex cover in the original graph. The tight relation between $\log L(S|T_{ML})$ and $\log L(S|T_{CA})$ implies a tight relationship between the size of an approximate vertex cover in the original graph and the maximum likelihood tree on the given sequences, and establishes the NP hardness of maximum likelihood on phylogenetic trees.

3 Model, Definitions and Notations

In this section we describe the model and basic definitions that we will use later. These definitions include phylogenetic trees and characters, the parsimony score, Neyman's two state model, and the likelihood function. In most of this paper, we assume that characters are in one of two states, 0 or 1. Let $S = [s(1), s(2), s(3), \ldots, s(n)] \in \{0,1\}^{n \times k}$ be the observed sequences of length k over n taxa (n leaves). Given such sequences, both the maximum parsimony and the maximum likelihood criteria aim at finding the tree (or trees) that "best explain" this data. Each uses a different objective function. In this section, both are defined and explained.

Definition 3.1 *(Phylogenetic trees, characters, labelings [22])*
A phylogenetic tree with n leaves is a tree $T = (V(T), E(T))$ such that each leaf (degree one vertex) is given a unique label from $[n] = \{1, ..., n\}$. A non leaf vertex is called an internal vertex. A function $\lambda : [n] \to \{0, 1\}$ is called a state function for T. A function $\hat{\lambda} : V(T) \to \{0, 1\}$ is called an extension of λ on T if it coincides with λ on the leaves of T. In a similar way, we define a function $\lambda^k : [n] \longmapsto \{0, 1\}^k$ and an extension $\hat{\lambda}^k : V(T) \longmapsto \{0, 1\}^k$. This later function is called a labelling of T. If $\hat{\lambda}^k(v) = s$ we say that the string s is the labelling of the vertex v.
Given a labelling $\hat{\lambda}^k$, let $d_e(\hat{\lambda}^k)$ denote the number of differences between two the labellings of the endpoints of the edge $e \in E(T)$.

Definition 3.2 *(Maximum parsimony score)*
Let T be a tree with n leaves, and S be a set of n binary strings, all of length k. Let $\lambda^k{}_{pars} : [n] \longmapsto S$ be labeling of T's leaves: An mapping onto the strings S. Let $\hat{\lambda}^k_{pars} : V(T) \longmapsto \{0, 1\}^k$ be an extension of λ^k that minimizes the expression $\sum_{e \in E(T)} d_e(\hat{\lambda}^k)$. We define $pars(S, T, \lambda^k)$, the parsimony score for S, T, λ^k, as the value of this sum. A maximum parsimony tree (or trees) for the set of binary strings, S, is a tree (or trees) and leaf labeling that minimizes the sum above over all trees T and assignments λ^k of the strings in S to leaves' labelings. The value of the sum on this tree is called the parsimony score for the set of strings S.

When the labeling $\hat{\lambda}^k$ is clear, we simply use d_e instead of $d_e(\hat{\lambda}^k)$. In the likelihood setting, we endow edges with "mutation probabilities". For a tree T, let $\mathbf{p} = [p_e]_{e \in E(T)}$ be the edge probabilities. We use the Neyman two states model [18]. Given labels of length k, each position $j \in \{1, \ldots, k\}$ is called a *site*. According to this model:

- Leaves' labels are strings from $\{0, 1\}^k$.
- The "edge probability" p_e satisfies $0 \leq p_e \leq \frac{1}{2}$.
- The probability of a net change of state (from '1' to '0' or vice versa) occurring across an edge e (a "mutation event") is given by p_e. This probability is also called the "length", or "weight", of edge e.
- Mutation (change) events on different edges are independent.
- Different sites mutate independently.

The likelihood of observing an $S \in \{0, 1\}^{n \times k}$, given the tree T with $r \leq n - 2$ internal nodes and the edge probabilities $\mathbf{p}$, $L(S|T, \mathbf{p})$, is defined as

$$L(S|T, \mathbf{p}) = \prod_{i=1}^{k} \sum_{\mathbf{a} \in \{0,1\}^r} \prod_{e \in E(T)} m(p_e, S_i, a_i) \ , \tag{1}$$

where $\mathbf{a}$ ranges over all combinations of assigning labels (length k 0 or 1 strings) to the r internal nodes of T. This notion of ML is termed maximum *average* likelihood in Steel and Penny [21]. Each term $m(p_e, S_i, a_i)$ is either p_e or $(1-p_e)$, depending on whether in the i-th site of S and $\mathbf{a}$, the two endpoints of e are assigned different characters states (and then $m(p_e, S_i, a_i) = p_e$) or the same characters states (and then $m(p_e, S_i, a_i) = 1 - p_e$). The ML solution (or solutions) for a specific tree T is the point (or points) in the edge space $\mathbf{p} = [p_e]_{e \in E(T)}$ that maximizes the expression $L(S|T, \mathbf{p})$. The global ML solution is the pair (or pairs) $(T, \mathbf{p})$, maximizing the likelihood over all trees T of n leaves, labeled by S, and all edge probabilities $\mathbf{p}$ (for more details see [8, 20, 22]). By

the independence of sites, an equivalent way to define the likelihood of observing S in the tree T is:

$$L(S|T,\mathbf{p}) = \sum_{\lambda^k \in \{0,1\}^{k\times r}} \prod_{e\in E(T)} p_e^{d_e(\lambda^k)} \cdot (1-p_e(\lambda^k))^{k-d_e(\lambda^k)} \tag{2}$$

In the rest of the paper we use this definition for likelihood.

4 Properties of Maximum Likelihood Trees

In this section we prove some useful properties of ML trees. We start with properties of general trees and continue with canonical ones.

4.0.1 General Properties of ML Trees

In our NP-hardness proof we want to show that the ML tree for a set of reduction strings, defined in the next section, have log likelihood arbitrarily close to the log likelihood of some canonical tree. We achieve this by a sequence of pruning sub-forests that satisfy certain conditions, and rearranging them in a canonical way around the "root". We will show bound on the decrease in the log likelihood by such rearrangements. The following Lemma is used several times in the rest of this paper.

Lemma 4.1 *Let T be a phylogenetic tree with edge probabilities $\mathbf{p}$, let S (set of binary string of length k) denote the labelling for the leaves of the tree. Suppose F_1 and F_2 are two disjoint forests that partition T, and have the node x as their common root. Let S_1 and S_2 be the leaf labellings of F_1 and F_2, respectively and let $\mathbf{p_1}$, $\mathbf{p_2}$ be the induced, corresponding edge probabilities. Let ℓ_x denote a labelling of x. Then the likelihood of observing S given T and $\mathbf{p}$ equals*

$$L(S|T,\mathbf{p}) = \sum_{s\in\{0,1\}^k} L(S_1, \ell_x = s|F_1, \mathbf{p_1}) \cdot L(S_2, \ell_x = s|F_2, \mathbf{p_2}).$$

Proof. Follows directly from equation (2). ■

For "standard" phylogenetic trees, the internal nodes do not have any specified labelling, while leaves are labelled by a k long sequence. In the course of our modifications we could also have a leaf with no labelling (see figure 2). The natural way to define the likelihood of a tree with such leaves is to treat them as internal nodes, namely summing over all their possible labellings. The next Lemma states that such "unlabelled" leaves can be pruned without effecting the likelihood.

Lemma 4.2 *Let T be a phylogenetic tree with an unlabelled leaf. By pruning this leaf (and the edge connecting to it) we get a tree, T', with equal likelihood.*

Proof. Let $S = \{S_i\}$ be the set of leaves' labels (binary strings of length k). Let h be the unlabelled leaf, and let h' be its neighbor in T. According to the definition:

$$\begin{aligned}
Pr(S \mid T,\mathbf{p}) &= \sum_{s\in\{0,1\}^k} Pr(S, \ell_h = s|T,\mathbf{p}) \\
&= \sum_{s\in\{0,1\}^k} \sum_{r\in\{0,1\}^k} Pr(S, \ell_{h'} = r|T',\mathbf{p}) \cdot Pr(\ell_h = s, \ell_{h'} = r|\mathbf{p}) \\
&= \sum_{r\in\{0,1\}^k} Pr(S, \ell_{h'} = r|T',\mathbf{p}) \\
&= Pr(S|T',\mathbf{p})
\end{aligned}$$

since $\sum_{s\in\{0,1\}^k} Pr(\ell_h = s, \ell_{h'} = r|\mathbf{p}) = 1$. ■

Lemma 4.3 *Let T be a phylogenetic tree with an internal node, h, of degree two, and let g_1, g_2 be its neighbors. Then h can be eliminated to create an (g_1, g_2) edge without changing the likelihood of T.*

Proof. Let p_{h,g_1} and p_{h,g_2} be the mutation probabilities of the edges (h, g_1) and (h, g_2), respectively. Set $p_{g_1,g_2} = p_{h,g_1}(1 - p_{h,g_2}) + p_{h,g_2}(1 - p_{h,g_1})$. It is easy to see that for $0 \leq p_{h,g_1}, p_{h,g_2} \leq 1$, we get $0 \leq p_{g_1,g_2} \leq 1$, and that the mutation probability across the path from g_1 to g_2 does not change. ■

Our NP completeness proof heavily uses a sequence of tree modifications. In each modification, we uproot a forest (a collection of subtrees with a common ancestor), rearrange it, and graft it on the root of the tree. The following theorem establishes a connection between the likelihood of the original and the rearranged forest, and the change in the total likelihood of the tree.

Theorem 4.4 *Let T be a phylogenetic tree, with edge probabilities $\mathbf{p}$, S a set of labels to its leaves, such that one of its leaves is labeled by the all zero sequence. Let* root *denote an internal node on T that is at distance 0 from this leaf. Suppose $T_1, .., T_j$ are a subforest of T, namely a collection of disjoint subtrees that have a common root, h. Denote by T^- the original subforest, let $\mathbf{p}^-$ be its edge probabilities, and S^- the labels at its leaves. Suppose we uproot T^-, rearrange it to a new subforest, T_{new}, and endow the new subforest with edge probabilities $\mathbf{p}_{new}$. Let $T_{arranged}$ be the tree resulting from grafting T_{new} onto root (the root of the forest T_{new} is the node root), where the edge probabilities $\mathbf{p}_{arranged}$ in the "old" part are as in $\mathbf{p}$. Suppose there is $W > 0$ such that for every labelling s of h,*

$$Pr(S^-, \ell_h = 0|T_{new}, \mathbf{p}_{new}) \ \geq W \cdot Pr(S^-, \ell_h = s|T^-, \mathbf{p}^-) \ .$$

then

$$Pr(S|T_{arranged}, \mathbf{p}_{arranged}) \ \geq W \cdot Pr(S|T, \mathbf{p}).$$

Proof. The likelihood of S, given the initial tree, T, and $\mathbf{p}$, is (see Lemma 4.1):

$$L_1 \equiv L(S|T, \mathbf{p}) = Pr(S|T, \mathbf{p}) = \sum_{s\in\{0,1\}^k} Pr(S^-, \ell_h = s|T^-, \mathbf{p}^-) \cdot Pr(S \setminus S^-, \ell_h = s|T \setminus T^-, \mathbf{p} \setminus \mathbf{p}^-) \ .$$

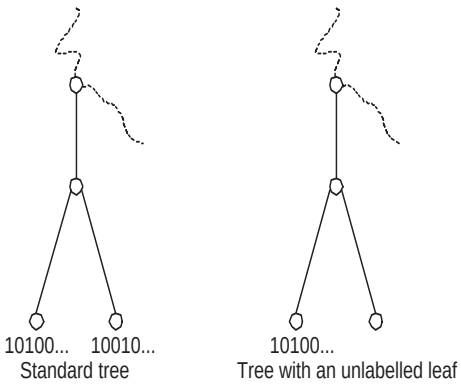

Figure 2: Standard and nonstandard trees

The likelihood of S given $T_{arranged}$ and $\mathbf{p}_{arranged}$ equals

$$L_2 \equiv L(S|T_{arranged}, \mathbf{p}_{arranged}) = Pr(S|T_{arranged}, \mathbf{p}_{arranged})$$
$$= \sum_{s\in\{0,1\}^k} Pr(S^-, \ell_h = 0^k|T_{new}, \mathbf{p}_{new}) \cdot Pr(S \setminus S^-, \ell_h = s|T_{arranged} \setminus T^-, \mathbf{p}_{arranged} \setminus \mathbf{p}_{new}) \ .$$

According to our assumption, for each $s \in \{0,1\}^k$,

$$Pr(S^-, \ell_h = 0^k|T_{new}, \mathbf{p}_{new}) \geq W \cdot Pr(S^-, \ell_h = s|T^-, \mathbf{p}^-) \ ,$$

and thus $L_2 \geq W \cdot L_1$, as desired. ■

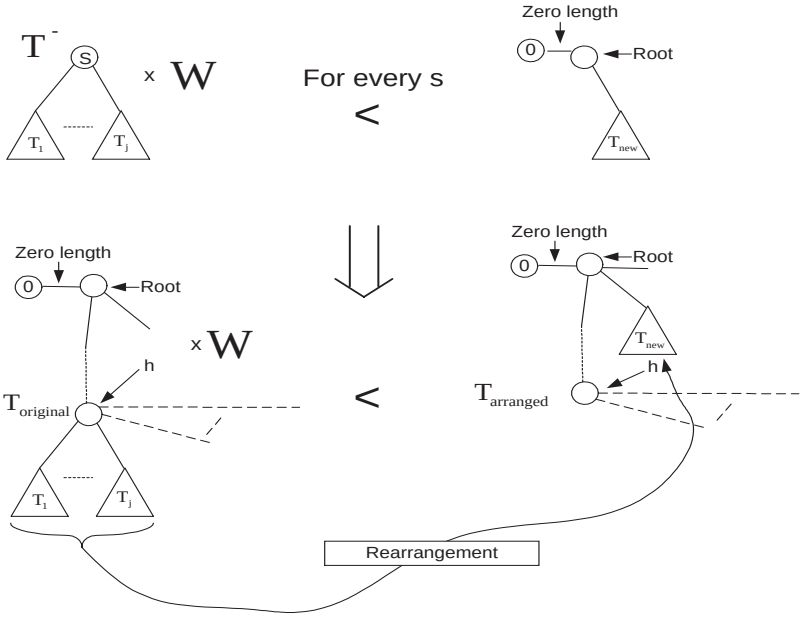

Figure 3: Theorem 4.4.

The following corollary is a direct result of Theorem 4.4, and uses the same notation.

Corollary 1 *Let S^- denote the strings at the leaves of T^-, and let s denote a labelling of h. Let $T^*(s)$, $\mathbf{p}^*(s)$, respectively, denote the structure and edge length of a tree that maximize the likelihood of the strings $S^- \cup \{s\}$. Let s^* be a string that maximizes this likelihood, namely for all $s \in \{0,1\}^k$, $Pr(S^- \cup \{s\}|T^*(s), \mathbf{p}^*(s)) \leq Pr(S^- \cup \{s^*\}|T^*(s^*), \mathbf{p}^*(s^*))$. If $Pr(S^-, \ell_h = 0^k|T_{new}, \mathbf{p}_{new}) \geq W \cdot Pr(S^- \cup \{s^*\}|T^*(s^*), \mathbf{p}^*(s^*))$, then $Pr(S|T_{arranged}, \mathbf{p}_{arranged}) \geq W \cdot Pr(S|T, \mathbf{p})$.*

The proof of the following lemma is immediate.

Lemma 4.5 *Let T^S_{ML}, $\mathbf{p}$ denote the structure and edges lengths of an ML tree for the set of strings S. For any $s \in \{0,1\}^k$, $Pr(S \cup \{s\}|T^{S\cup\{s\}}_{ML}, \mathbf{p}') \leq Pr(S|T^S_{ML}, \mathbf{p})$, where $T^{S\cup\{s\}}_{ML}$ and $\mathbf{p}'$ are optimal structure and edge lengths for $S \cup \{s\}$, respectively.*

In the following lemma we are interested in the subforest T^-, and ignore the rest of the tree. We are allowed to change the structure of T^-, and want to find a labelling s^* of its root, h, such that the likelihood of $S^- \cup \{s^*\}$, given the new structure of T^- is maximized. The maximization of the likelihood is over all the labelling of h, and over all structures for the subforest. We denote such a labelling s^* of h an optimal root labelling.

Lemma 4.6 *Suppose there are sites where the value of all the strings in S^- is 0. Then there is an optimal root labelling s^* whose value in any of these sites is also 0.*

Proof. Let $T_{ML}^{S^-}$, $\mathbf{p}^-$ denote the structure and edges lengths of an ML tree for the set of strings S^-. Take any of the leaves in this tree, labeled by some $s_i \in S^-$. Make this leaf into two leaves, connected by a pair of 0 length, and label the two leaves by s_i as well. It is not hard to see that the new tree has exactly the same likelihood as $T_{ML}^{S^-}$, $\mathbf{p}^-$. By lemma 4.5 for any $s \in \{0,1\}^k$, $Pr(S^- \cup \{s\}|T_{ML}^{S^- \cup \{s\}}, \mathbf{p}') \leq Pr(S^-|T_{ML}^{S}, \mathbf{p}^-)$. Since for $s = s_i$ we have equality, it follows that any original string is also an optimal root labeling. In particular, there are optimal root labeling that have a 0 in any site where all strings in S^- have a 0. ■

For any tree T on m leaves and any observed sequences S, we denote by $\mathbf{p}^*$ the edge probabilities that maximize $L(S|T, \mathbf{p})$. The following Theorem is a restatement of Theorem 7 from Tuffley and Steel [22].

Theorem 4.7 *Let S be a set of m binary strings of length $k = k_c + k_{nc}$, where k_c is the number of constant characters in S (i.e. positions that have the same value for all the strings). Let T be a tree on m leaves. Let $pars(S,T)$ denote the parsimony score of S on the tree T. Then*

$$2^{-\log(k_c)\cdot pars(S,T) - C^d_{T,pars(S,T)}} \leq L(S|T) \triangleq Pr(S|T,\mathbf{p}^*) \leq 2^{-\log(k_c)\cdot pars(S,T) - C^u_{T,pars(S,T)}}$$

and

$$\lim_{k_c \to \infty} \frac{-\log(Pr(S|T,\mathbf{p}^*))}{\log(k_c)} = pars(S,T)$$

where $C^u_{T,pars(S,T)}, C^d_{T,pars(S,T)} = O(k_{nc}m + pars(S,T)\log pars(S,T))$.

If we hold m fixed and pad the strings in S, then k_c increases, but $pars(S,T)$ remains invariant. The first terms in the exponents of both the upper and lower bounds become dominant. This establishes the limit, and furthermore the fact that the ML tree "converges" to an MP tree.

Corollary 2 *Let S contain m binary sequences of length k. Let T_a and T_b be two trees with he strings of S in their leaves. Let $\mathbf{p}_a^*$ and $\mathbf{p}_b^*$ denote the optimal edges' length for S on these two trees, respectively. Suppose that the strings in S have k_c many constant sites. Then there is k_c large enough (it at least should be $2^{C^u_{T,pars(S,T)}} = o(k_c)$) such that $pars(S,T_a) < pars(S,T_b)$ implies $Pr(S|\mathbf{p}_a^*, T_a) > Pr(S|\mathbf{p}_b^*, T_b)$.*

We remark that in general equality in the parsimony score does not imply equality in the likelihood. The next corollary generalizes the previous one to general trees with one internal node that is labelled. (The definition of likelihood for such trees is straightforward, and is omitted.) Suppose S is a set containing length k strings, which share k_c constant positions. The likelihood, $Pr(S, \ell_h = s|F, \mathbf{p}^*)$, of a subforest F with r subtrees $T_1, .., T_r$, sets of labellings of the subtrees' leaves $S_1, .., S_r$, optimal edges' lengths $\mathbf{p}^*$, and with a label $\ell_h = s$ at the root of the subforest (see figure 3) is

$$Pr(S, \ell_h = s|F, \mathbf{p}^*) = \prod_{i=1}^{r} Pr(S_i, \ell_h = s|\mathbf{p}^*, T^i) \ .$$

Therefore

$$Pr(S, \ell_h = s|F, \mathbf{p}^*) \geq \prod_{i=1}^{r} 2^{-\log(k_c)\cdot pars(S_i \cup \{s\}, T_i) - C^d_{T_i, pars(S_i \cup \{s\}, T_i)}} \ , \text{ and}$$

$$Pr(S, \ell_h = s|F, \mathbf{p}^*) \leq \prod_{i=1}^{r} 2^{-\log(k_c)\cdot pars(S_i\cup\{s\},T_i) - C^u_{T_i,pars(S_i\cup\{s\},T_i)}} ,$$

where $C^u_{T_i,pars(S_i\cup\{s\},T_i)}$ and $C^d_{T_i,pars(S_i\cup\{s\},T_i)}$ are the functions defined in Theorem 4.7.
Let $pars(S \cup \{s\}, F) = \sum_i pars(S_i \cup \{s\}, T_i)$, and $C^u_{S\cup\{s\},F} = \sum_i C^u_{T^i,pars(S_i\cup\{s\},T_i)}$ and let $C^d_{S\cup\{s\},F} = \sum_i C^d_{T^i,pars(S_i\cup\{s\},T_i)}$. Summing up the exponents, we get

Corollary 3

$$2^{-\log(k_c)\cdot pars(S\cup\{s\},F) - C^d_{S\cup\{s\},F}} \leq Pr(S, \ell_h = s|F, \mathbf{p}^*) \leq 2^{-\log(k_c)\cdot pars(S\cup\{s\},F) - C^u_{S\cup\{s\},F}} .$$

Proof. The proof follows directly from Theorem 4.7 and the properties of our model. ■

Let S denote the set of labelling of the leaves of a forest $F = (V, E)$. Let k_{nc} denote the number of non constant sites in $S \cup \{s\}$, where $\ell_h = s$ is the labelling at the F's root, h. The "reduction strings" we will deal with have the property that the number of non-constant sights of the labellings of the forest F is small, namely $k_{nc} \leq 2|F|$.

In this case by theorem 4.7 and corollary 3 we get the following relationship between the log likelihood and the parsimony:

$$\log L(S, \ell_h = s|F, \mathbf{p}^*) = O(pars(S, F) \cdot \log(k_c)) + O(pars(S, F) \cdot \log(pars(S, F))) + O(|F|^2)$$

4.0.2 Properties of Canonical ML Trees

In this subsection, we study properties related to canonical ML trees (definition 2.1), properties that play an important role in our reduction. Throughout this section, the strings we deal with are binary "reduction strings" of length n, originating from a graph of n nodes and m edges.

Definition 4.8 *Let T_{C_i} ($i = 1,2,$ or 3) be a tree with $i+1$ leaves and one internal node (*i.e. *T_{C_i} has the star topology), such that one of the strings in the leaves is the all zero string (of length $k = n$). The other i strings are all of weight 2 (two 1s), and for $i > 1$ they all share one "1" position (see figure 4). Let $ML_i(n)$ be the log ML score of T_{C_i} for the optimal edges' length of the tree. Let S_{C_i} denote the strings in the leaves of tree T_{C_i}.*

It is easy to see that $ML_i(n)$ does not depend on the specific choice of strings in T_{C_i}.

Lemma 4.9 *There are constants $C^d \leq C^u$ such that for all for n large enough, the following properties hold:*

1. $-2 \cdot \log(n) + C^d \leq ML_1(n) \leq -2 \cdot \log(n) + C^u$
2. $-3 \cdot \log(n) + C^d \leq ML_2(n) \leq -3 \cdot \log(n) + C^u$
3. $-4 \cdot \log(n) + C^d \leq ML_3(n) \leq -4 \cdot \log(n) + C^u$

Proof. The proof follows from Theorem 4.7 and direct calculations. ■

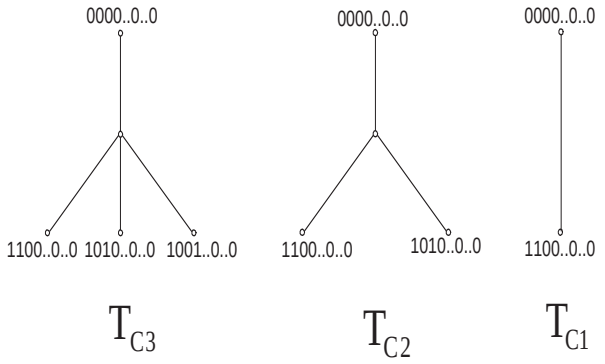

Figure 4: Building blocks of the maximum likelihood tree for our reduction.

Theorem 4.10 *Let S be a set of "reduction strings", corresponding to a graph with n nodes and m edges. Let T be a canonical trees with $m+1$ leaves, labelled by S. Let d denote the degree of its root. Let $\mathbf{p}^*$ be optimal edges weights for S with respect to this tree. Then for the constants $C^d \leq C^u$ of Lemma 4.9,*

$$-(d+m)\cdot\log n + dC^d \leq \log Pr(S|T,\mathbf{p}^*) \leq -(d+m)\cdot\log n + dC^u \ .$$

Proof. According to Lemma 4.9, $ML_i(n)$, the log likelihood of a subtree T_{c_i} with i non-zero leaves and n long labellings, that is hung off the root of the canonical tree, satisfies

$$-(i+1)\log(n) + C^d \leq ML_i(n) \leq -(i+1)\log(n) + C^u \ .$$

Since T's root is effectively labeled by the all zero vector, the log likelihood of S given T is obtained by summing all log likelihoods of its d subtrees. All subtrees together have m leaves, other than the all zero leaf. So the log likelihood of S given T satisfies

$$-(d+m)\log(n) + dC^d \leq \log L(S|T,\mathbf{p}^*) \leq -(d+m)\log(n) + dC^u \ .$$

■

Let S be a set of "reduction strings", corresponding to a graph with n nodes and m edges. Let T_a and T_b be two canonical trees with $m+1$ leaves, labelled by S. Let d_a and d_b denote the degrees of the roots of these trees, correspondingly. Let p_a^* and p_b^* be optimal edges weights for S with respect to these trees. Then for the constants $C^d \leq C^u$ of Lemma 4.9, the log likelihood ratio of these trees satisfies

$$\frac{-(d_a+m)\cdot\log n + d_a\cdot C^d}{-(d_b+m)\cdot\log n + d_b\cdot C^u} \leq \frac{\log Pr(S|T_a,\mathbf{p}_a^*)}{\log Pr(S|T_b,\mathbf{p}_b^*)} \leq \frac{-(d_a+m)\cdot\log n + d_a\cdot C^u}{-(d_b+m)\cdot\log n + d_b\cdot C^d} \ .$$

Since $d_a, d_b \leq n$, both the left hand side and the right hand side converge to $(d_a+m)/(d_b+m)$ as n grows. This implies that for large enough n,

Corollary 4 *For any arbitrarily small ε there is n_0 large enough such that for all $n \geq n_0$*

$$\frac{d_a+m}{d_b+m}\cdot(1-\varepsilon) \leq \frac{\log Pr(S|p_a^*,T_a)}{\log Pr(S|p_b^*,T_b))} \leq \frac{d_a+m}{d_b+m}\cdot(1+\varepsilon).$$

And in particular if $d_a = d_b$, then

$$lim_{n\to\infty}\frac{\log Pr(S|T_a, p_a^*)}{\log Pr(S|T_b, p_b^*))} = 1.$$

5 NP-Hardness of Maximum Likelihood

Building upon the ML machinery developed so far, we now turn to the proof that ML reconstruction on trees is NP hard. We start by formally defining the decision version of maximum likelihood, and then of the gap version of vertex cover we use.

Problem 5.1 *Maximum likelihood (ML).*
Input: *S, A set of binary strings, all of the same length, and a negative number L.*
Question: *Is there a tree, T, such that* $\log Pr(S|T, \mathbf{p}^*(S,T)) > L$ *?*

A gap vertex cover problem is the following:

Problem 5.2 *Gap problem for vertex cover, $gap - VC[c_1, c_2]$*
Input: *A graph, $G = (V, E)$, two positive numbers, c_1 and c_2.*
Task: *Does G have a vertex cover smaller than c_1, or is the size of each vertex cover is larger than c_2 ? (If the minimum vertex cover is in the intermediate range, there is no requirement.)*

Our proof emplys a reduction from the gap version of vertex cover, restricted to degree 3 graphs (undirected graph of degree at most 3 in each node). We rely on the following hardness result of Karpinski and Berman [2].

Theorem 5.3 *[2] The following problem* [1]*, $gap - VC_3[\frac{144}{284}\cdot n, \frac{145}{284}\cdot n]$, is NP-hard: Given a degree 3 graph, G, on n nodes, is the minimum VC of G smaller than $\frac{144}{284}\cdot n$, or is it larger than $\frac{145}{284}\cdot n$?*

We reduce that specific version of $gap - VC_3$ above to ML.

5.1 Reduction and Proof Outline

Given an instance $G = (V, E)$ of $gap - VC_3$, denote $|V| = n$, $|E| = m$, $c_1 = \frac{144}{284}\cdot n$ and $c_2 = \frac{145}{284}\cdot n$. We construct an instance $< S, L >$ of ML such that S is a set of $m + 1$ strings, each string of length $k = n$, and $L = -(m + \frac{c_1+c_2}{2}) \cdot \log n$.

The first string in S consists of all zeros (the all zeros string), $\underbrace{00...0...00}_{k}$, and for every edge $e = (i, j) \in E$ there is a string, $S(e) = \underbrace{\overbrace{00..00}^{i-1}1\overbrace{00..00}^{j-i-1}1\overbrace{00..00}^{k-j}}_{k}$ where the i-th and the j-th positions are set to 1, and all the rest are set to 0. These m strings are called "edge strings". From now on, the trees we refer to have leaves whose labels are generated by this construction.

We use asymptotic properties of likelihood of trees, so most claims will hold when the input graph is large enough (*i.e.* $n = |V|$ is large enough). In our proof, we deal with small size subtrees or forests, containing at most $2 \cdot \log\log n$ leaves.

[1] We could also use the deep gap VC results of Håstad [14]) and Dinur and Sufra [7]. However their graphs are of bounded degree greater than 3 and it seems that the modification to bounded degree 3 graphs would yield smaller gaps (not effecting the hardness of ML, though).

We will need the following relation for the expressions in the likelihood of the forests to hold (see corollary 3):

$$\frac{C^d_{S\cup\{s\},F}}{\log(k_c)\cdot pars(S\cup\{s\},F)}, \frac{C^u_{S\cup\{s\},F}}{\log(k_c)\cdot pars(S\cup\{s\},F)} \rightarrow_{n\rightarrow\infty} 0$$

By lemma 4.6 we can assume s have 0 in positions were all the forest's strings have 0. The parsimony score (and k_{nc}, the number of non-constant sites) of such forest is no more than $4\cdot\log\log n$, thus $C^d_{S\cup\{s\},F}, C^u_{S\cup\{s\},F} = O((\log\log n)^2)$. Since $k_c = O(n)$ we can use the quantitative relations between parsimony and likelihood as proved in corollary 3, our proof is strongly relies on these relations.

5.2 From ML to Canonical Trees

In this section we show that for every $\varepsilon > 0$, there is an $n_0 > 0$ such that for $n > n_0$, the ratio between the log likelihood and the maximum log likelihood of some canonical tree is upper bounded by $(1+\varepsilon)$.

Given an ML tree, T, if it is in canonical form, we are done. Otherwise we locate subtrees of T, $T_1, T_2, ..., T_\ell$ with a common root, such that the number of leaves in $\bigcup_{i=1}^{\ell} T_i$ is in the interval $[\log\log n, 2\cdot\log\log n]$. Notice that this is a forest as there may be other subtrees rooted at the same node. It is easy to show that such a forest always exists (lemma 5.4).

Lemma 5.4 *Suppose T is a rooted tree and v is an internal node such that the number of leaves below v is at least q. Then v has a descendent, u, such that u is the root of a forest consisting of ℓ subtrees $T_1, T_2, ..., T_\ell$ ($\ell \geq 1$), and the number of leaves in the forest $\bigcup_{i=1}^{\ell} T_i$ is in the range $[q, 2\cdot q]$.*

The next lemmata, we show that the ratio of the log-likelihood of such forest when the all zero labelling is placed in its root, and the log-likelihood of the same subforest with the best labelling in its root, is close to 1.

Lemma 5.5 *Let u be an internal node or the root h in F, whose degree is $r \geq 9$, and let $s \in \{0,1\}^k$. Consider an assignment of labels to internal nodes of F, where h is assigned s. Among such assignments, those that optimize the parsimony score label u with 0^k.*

Proof. We assume here $u \neq h$, the case were $u = h$ can be proved in a similar way. It suffices to prove the claim for every position separately. The internal node u have $r-1$ subtrees below it, and one edge "above" it, leading to h. Out of these subtrees, at most three have "1" in the position of interest (since our graphs are of degree 3). For the other $r-4 > 4$ subtrees, since their leaves have 0 in the position, the most parsimonious assignment will label all their nodes with 0, as can be seen by running Fitch algorithm [11]. Therefore u has at least five neighbor nodes with 0 in this position, and at most four with 1. Any parsimonious assignment will thus label u with 0. ■

Lemma 5.6 *Let h be the root of a forest F (h has at least two children in F) whose leaves are labelled by reduction strings S. Suppose that in each position, the leaves labelled with "1" are at distance at least 4 from h. Then the max parsimony score for S on F is achievable with the all zero labelling in h.*

Proof. Consider an arbitrary position. Since the reduction strings emanate from a degree 3 graph, there are at most three leaves x,y,z with '1' in this position. Let $LCA(x,y)$, $LCA(x,y,z)$

denote the least common ancestors of x, y and x, y, z, respectively (see figure 5). Suppose, without loss of generality, that $LCA(x, y)$ is equal to $LCA(x, y, z)$ or is below it in F. For any node j in F, we denote by $pa(j)$ the parent of j. There are three cases:

1. $h = LCA(x, y, z)$ and $LCA(x, y) \neq LCA(x, y, z)$: Consider the path from z to h. At most one feeding node, the one leading to it from $LCA(x, y)$, may be assigned '1' in the a optimal parsimony assignment. There are at least two other nodes feeding to the path. Therefore, by Fitch algorithm, the best assignment to the nodes in the path from z to h is '0'. Thus if we assign '0' to h we may lose 1 in the score due to the node just below h in the path between $LCA(x, y)$ and h, but lose nothing in the score due to the edges to the other children of h. On the other hand, if we assign '1' to h we lose 1 in the score due to the node before last in the path from z to h, and may lose 1 in the score due to the node just below h in the path between $LCA(x, y)$ and h. Thus the '0' labelling to h is not worse than the '1' labelling.

2. $h = LCA(x, y, z)$ and $LCA(x, y) = LCA(x, y, z)$: By Fitch algorithm, the best assignment to the nodes in the paths from z, y, and x to h is '0'. If we assign '1' to h we loose 1 on each edge leading to h, a total loss of 3. If we assign '0' to h we lose nothing on the node immediately below to h.

3. $h \neq LCA(x, y, z)$:

   Since h has at least two children, all the leaves under one of these children are '0', so the algorithm of Fitch assigns '0' to this node. Since $LCA(x, y, z)$ is below h, Fitch's algorithm will assign "1" to at most one of h children. Thus the '0' labelling to h is not worse than the '1' labelling.

■

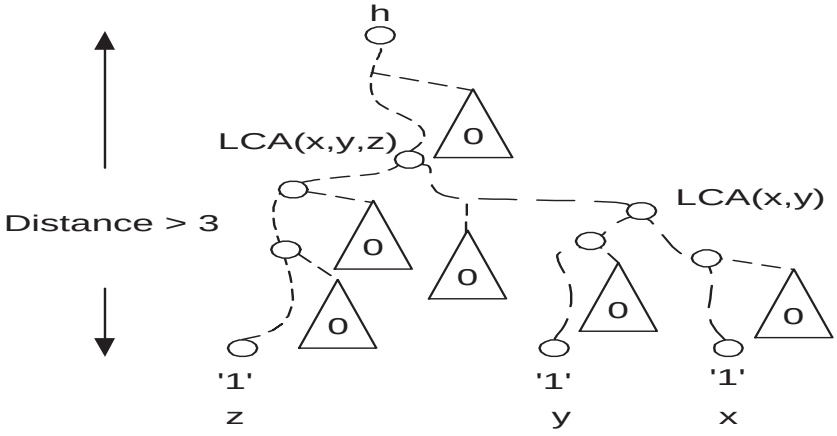

Figure 5: Lemma 5.6, case (3).

Corollary 5 *Let h be a root of a subforest F whose leaves are labelled by a set of reduction strings, S. Consider a specific position, and suppose all leaves having "1" in the position are either at distance ≥ 4 from the root, or have an internal node of degree ≥ 9 on the path to the root. Then the parsimony score of F when labelling the root h with 0 at this position is at least as good as when labelling the root with 1.*

Proof. According to Lemma 5.5, if there is an internal node of degree ≥ 9, it is better to assign 0 to this node. This enables us to disregard the "1" leaves of distance smaller than 4 from h with a high degree node on their path to h. The other leaves with "1" are at distance at least 4 from h, and we can now apply Lemma 5.6. ■

Theorem 5.7 *Let T be a tree whose leaves are labelled by a subset of non-zero reduction strings. Let F be a subforest of T, rooted at h, and let $s \in \{0,1\}^k$ be a label of h. The parsimony score of F with the 0^k labelling at the root can be worse by at most $8^4 - 1$ than the parsimony score of F with label s at its root.*

Proof. If the degree of h is larger than 8, then by Lemma 5.5 the best assignment to h is 0^k and we are done. Otherwise, we say that a leaf in F is *dangerous* if its distance from h less than 4, and all its ancestors on the path to h have degree ≤ 8. Simple counting shows that the number of dangerous leaves in F is smaller than $8 + 8^2 + 8^3$. Every dangerous leaf has 2 positions where it is labelled "1". Each such position can be "1" in at most 3 leaves because the graph G is of degree 3. Therefore for any of these positions, changing the label at h from 1 to 0 will worsen the parsimony score by at most 3. There are at most $2 \cdot (8 + 8^2 + 8^3)$ such positions. So changing to 0 at h will cause at most $2 \cdot 3 \cdot (8 + 8^2 + 8^3) < 8^4$ parsimony degradations. According to Lemma 5, in all other positions, the "0" label at h is optimal. ■

The following Lemma was proved by Day *at. al.* [5, 1].

Lemma 5.8 *Let $S' \subseteq S$ be a subset of the reduction strings, which contains the all zero string. The structure of the best parsimony tree for S' is canonical.*

Theorem 5.9 *For every $\varepsilon > 0$ there is an n_0 such that for all $n \geq n_0$, if $S \subseteq \{0,1\}^n$ is a set of $m+1$ reduction strings corresponding to a degree 3 graph with n nodes and m edges, then the following holds: Let T_{ML} denote an ML tree for S, and let $\mathbf{p}^*_{ML}$ and be an optimal edges length for this ML tree. Then there is a canonical tree for S, T_{CA}, with optimal edges length $\mathbf{p}^*_{CA}$, such that*

$$\frac{\log Pr(S|T_{ML}, \mathbf{p}^*_{ML})}{\log Pr(S|T_{CA}, \mathbf{p}^*_{Ca})} > (1-\varepsilon) \ .$$

Proof. We start from any ML tree, T_{ML}, and show how to transform it to a canonical tree, T_{CA}, with "close enough" log likelihood, in a sequence of up to $n/\log\log(n)$ steps. Each step involves a small, local change to the current tree. We identify a forest (disjoint subtrees with a common root), whose number of leaves is in the interval $[\log\log(n), 2\log\log(n)]$. By Lemma 5.4, if the root of the whole tree has a subtree with more than $\log\log(n)$ leaves, we can find such a forest. In such case, we first prune this forest, and then regraft it under the tree's root. Let S_F be the restriction of the set S to the labeling of F's leaves. By Theorem 5.7, the parsimony score of S_F on F, with the 0^n label at the root, is larger by at most $B \stackrel{\triangle}{=} 8^4$ than the parsimony score of of S_F on F with any $s \in \{0,1\}^n$ label at its root. We will show how this implies that for every s, the likelihood $Pr(S_F, \ell_h = s|F, \mathbf{p}^*_s)$ is not much larger than $Pr(S_F, \ell_h = 0^n|F, \mathbf{p}^*_s)$. To prove this, we use the results of Tuffley and Steel (corollary 3), the properties of our strings, and the small size of the forest.

Let s^*, F^* be a string and a forest structure, respectively, for which $Pr(S_F, \ell_h = s|F, \mathbf{p}^*_s)$ is maximized. By lemma 4.6, such an s^* that has 1s only in positions where some $s \in S_F$ has a 1 exists.

Clearly, instead of bounding the likelihood difference for every $s \in \{0,1\}^n$, it suffices to bound $Pr(S_F, \ell_h = s^*|F^*, \mathbf{p}^*_s) - Pr(S_F, \ell_h = 0^n|F, \mathbf{p}^*_s)$.

The parsimony score of $S_F \cup \{0^n\}$ on F, $pars(S_F \cup \{0^n\}, F)$, is no larger than the number of "1" entries in $S_F \cup \{0^n\}$. There are at most two "1"s per string, and the number of non-zero strings is at most $2 \log\log n$. Therefore $pars(S_F \cup \{0^n\}, F) \leq 4 \log\log n$. Since s^* can have a "1" only in positions where some $s \in S_F$ has a "1", the total number of "1"s in $S_F \cup \{s^*\}$ is at most twice the number in $S_F \cup \{0^n\}$. Therefore, $pars(S_F \cup \{s^*\}, F^*) \leq 8 \log\log n$.

Let us denote by k_{s^*} the number of constant sites with s^* at the root of F^*, and by k_0 the number of constant sites with 0^n at the root of F, by par_{s^*} the parsimony score of F^* with s^* at the root, and by par_0 the parsimony score of F with 0^n at the root. Then we shaw that $par_0 - par_{s^*} \leq B$, and know that k_{s^*}, k_0 are both $n - O(\log\log n)$, implying $\log(k_0) - \log(k_{s^*}) = \theta(1)$. Finally, by Theorem 4.7, the order of magnitude of both $C^u_{S_F \cup \{0^n\}, F}$ and $C^d_{S_F \cup \{s^*\}, F^*}$ is $O(k_{nc} m + \text{pars} \cdot \log \text{pars})$, where $k_{nc} = O(\log\log n)$ is the number of non-constant sites in the set of strings, $m \leq 2 \log\log n$ is the number of strings, and $\text{pars} = O(\log\log n)$ is the parsimony value. All by all, in our case $C^u_{S_F \cup \{0^n\}, F} - C^d_{S_F \cup \{s^*\}, F^*}$ are both $O((\log\log n)^2)$, and so is their difference. These inequalities imply

$$
\begin{aligned}
&\log\left(Pr(S_F, \ell_h = s^* | F^*, \mathbf{p}^*_{s^*})\right) - \log\left(Pr(S_F, \ell_h = 0^n | F, \mathbf{p}^*_0)\right) \\
&\leq -\log(k_{s^*}) \cdot par_{s^*} - C^u_{S_F \cup \{s^*\}, F^*} - \left(-\log(k_0) \cdot par_0 - C^d_{S_F \cup \{0^n\}, F}\right) \\
&= \log(k_0) \cdot (par_0 - par_{s^*}) + (\log(k_0) - \log(k_{s^*})) \cdot par_{s^*} + \left(C^u_{S_F \cup \{0^n\}, F} - C^d_{S_F \cup \{s^*\}, F^*}\right) \\
&\leq B \log(k_0) + \theta(par_{s^*}) + \left(C^u_{S_F \cup \{0^n\}, F} - C^d_{S_F \cup \{s^*\}, F^*}\right) \\
&\leq B \log(k_0) + O(\log\log(n)) + O((\log\log n)^2) \\
&\leq B \log(n) + o(\log n)
\end{aligned}
$$

To summarize, for each $s \in \{0,1\}^n$ we get

$$\log\left(Pr(S_F, \ell_h = s | F, \mathbf{p}^*_s)\right) - \log\left(Pr(S_F, \ell_h = 0^n | F, \mathbf{p}^*_0)\right) \leq B \log(n) + o(\log n) \ .$$

Let $T_{\text{arranged}}, \mathbf{p}^*_{\text{arranged}}$ denote the tree resulting from uprooting and regrafting the forest F, and its optimal edge length. The conditions of Theorem 4.4 apply, so we conclude

$$\log Pr(S | T, \mathbf{p}^*) - \log Pr(S | T_{\text{arranged}}, \mathbf{p}^*_{\text{arranged}}) \leq B \log(n) + o(\log n) \ ,$$

Namely each single uprooting decreases the overall log likelihood of S by no more than $B \log(n) + o(\log n)$. All uprootings therefore decrease the log likelihood by at most $Bn \log(n) / \log\log n + o(n \log n / \log\log n)$.

After a sequence of up to $n / \log\log n$ such uprootings, we get a tree having no subtrees with $\log\log n$ or more leaves. To get the desired canonical tree, we *separately* "canonize" each small subtree, namely rearange it in an optimal canonical form. According to Lemma 5.8, we can rearrange such a forest in a canonical form, with the all zero root, such that its parsimony score does not deteriorate. Let F_c denote such canonical rearrangement, and let $\mathbf{p}^*_c$ denote the optimal edges' length for the rearrangement, and k_c the number of constant sites in the strings set. By corollary 3,

$$
\begin{aligned}
& \log Pr(S, \ell_h = 0^n | F, \mathbf{p}^*_0) - \log Pr(S, \ell_h = 0^n | F_c, \mathbf{p}^*_c) \\
\leq\ & -\log(k_c) \cdot pars(S \cup 0^n, F) - C^u_{S \cup \{0^n\}, F} - \left(-\log(k_c) \cdot pars(S \cup 0^n, F_c) - C^d_{S \cup \{0^n\}, F_c}\right) \\
=\ & \log(k_c) \left(pars(S \cup 0^n, F_c) - pars(S \cup 0^n, F)\right) + \left(C^d_{S \cup \{0^n\}, F_c} - C^u_{S \cup \{0^n\}, F}\right)
\end{aligned}
$$

$$\begin{aligned} &\leq C^d_{S\cup\{0^n\},F_c} - C^u_{S\cup\{0^n\},F} \\ &= O((\log\log n)^2) \end{aligned}$$

Therefore, each such rearrangement can decrease the log likelihood of S given T by at most $O((\log\log n)^2)$. The minimal size of a forest that needs rearrangement is 2, so here are no more than $n/2$ forests to be rearranged. Overall, the decrease in log likelihood due to the rearrangements is $O(n(\log\log n)^2)$. Taking into both uprootings and rearrangements, the total log likelihood loss of the process is $Bn\log(n)/\log\log n + o(n\log n/\log\log n) + O(n(\log\log n)^2) = o(n\log n)$.

According to Theorem 4.10, the log-likelihood of all canonical trees is larger than $-n\log(n)$. We just showed the existence of a canonical tree whose log likelihood differs from the log likelihood of any ML tree by less than $Bn\log(n)/\log\log(n) + o(n\log(n)/\log\log(n))$. Thus there must be a constant $K > 0$ such that the log-likelihood of any ML tree is at most $-K\cdot n\log(n)$, and consequently there is a canonical tree such that the ratio between the log likelihood of the ML tree and this tree is

$$\frac{-K\cdot n\log(n)}{-K\cdot n\log(n) - O(n\cdot\log n/\log\log(n))} = 1 + O(\frac{1}{\log\log n})$$

implying that for every *varepsilon* > 0 there is an n_0 such that for all $n > n_0$
$L(S|T_{\mathrm{ML}})/L(S|T_{CA}) < 1+\varepsilon$. ■

We remark that the size of the subforests could be chosen to be different than $\theta(\log\log n)$ and still get similar result, provided they are neither too small nor too large.

5.3 Validity of the Reduction

In this section we complete the proof, by showing that indeed we have a reduction from $GAP{-}VC_3$ to ML. We show that if G has a small enough cover, then the likelihood of the corresponding canonical tree is high (this is the easy direction), and if the likelihood is high, then there is a small cover (this is the harder direction). The translation of sizes, from covers to log likelihood, and vice versa, is not sharp, but introduces some slack. This is why a gap version of vertex cover, instead of exact vertex cover, is required as our starting point.

The next Lemma establishes a connection between MP and VC, and was used in the NP-hardness proof of MP.

Lemma 5.10 *[5, 1] $G = (V, E)$ has a vertex cover of size c if and only if there is a canonical tree with parsimony score $c + m$, where c is the degree of the root.*

Theorem 5.11 *For every $0 < \varepsilon$ there is an n_0 such that for every $n \geq n_0$, if G is a degree 3 graph on n nodes and m edges, with a cover of size at most c, then there is a tree T such that the log - likelihood of S satisfies*

$$\log(Pr(S|T, \mathbf{p}^*(S,T))) > -(1+\varepsilon)(m+c)\log n.$$

On the other hand, if the size of every cover is $\geq c$, then the log likelihood of S given T satisfies

$$\log(Pr(S|T, \mathbf{p}^*(S,T))) < -(1-\varepsilon)(m+c)\log n.$$

Proof. Suppose G has a vertex cover of size $\leq c$. Since G's is of bounded degree 3, c satisfies $m/3 \leq c \leq m$, and $n \leq m \leq 1.5n$. According to Lemma 5.10, there is a canonical tree,

T_{CA}, with parsimony score $c+m$, such that the degree of its root is c. According to Theorem 4.10, the log likelihood of S, given this tree is, $-(c+m)\log(n)+\theta(n)$. Since $m, c = \theta(n)$, $\log(Pr(S|T_{CA}, \mathbf{p}^*(S,T_{CA}))) = -(c+m)\log(n)+\theta(n)$ implies that for every $\varepsilon > 0$ and large enough n,

$$\log(Pr(S|T_{CA}, \mathbf{p}^*(S,T_{CA}))) > -(m+c)\log(n)(1+\varepsilon) \ .$$

For the other direction, suppose the size of every cover of G is greater or equal to c. According to Lemma 5.10, the parsimony score of each canonical tree is at least $m+c$. Thus by Theorem 4.10, the likelihood of S, given any tree, is at most $-(m+c)\log(n)+cC^u$ (where C^u is the constant from the Theorem). Since $m, c = \theta(n)$ we get that for every $\varepsilon_1 > 0$ and large enough n,

$$-(m+c)\log(n)+cC^u < -(m+c)\log(n)(1-\varepsilon_1) \ .$$

According to Theorem 5.9, this implies that the likelihood of S with respect to any ML tree satisfies

$$\log(Pr(S|T_{ML}, \mathbf{p}^*(S,T_{ML}))) < -(m+c)\log(n)(1-\varepsilon_1)(1-\varepsilon_2) \ ,$$

where $\varepsilon_1, \varepsilon_2$ are arbitrarily small, and n is large enough. Thus, for every ε there is n_0 such that for $n > n_0$ the likelihood of the best trees satisfies

$$\log(Pr(S|T_{ML}, \mathbf{p}^*(S,T_{ML}))) < -(m+c)\log(n)(1-\varepsilon) \ .$$

■

Theorem 5.12 *ML reconstruction on trees is NP-hard.*

Proof. Let $G = (V, E)$ be an instance of $gap - VC_3$. Denote $|V| = n$, $|E| = m$, $c_1 = \frac{144}{284} \cdot n$ and $c_2 = \frac{145}{284} \cdot n$. Recall that in the reduction, we construct an instance $< S, L >$ of ML such that S is a set of $m+1$ strings, each string is of length $k = n$, and the threshold is $L = -(m + \frac{c_1+c_2}{2}) \cdot \log n$.

Suppose $G \in gap - VC_3$. Then G has a cover of size $\leq c_1$. According to Theorem 5.11, for every $\varepsilon > 0$ and large enough n, $\log(Pr(S|T_{ML}, \mathbf{p}^*(S,T_{ML}))) > -(m+c_1)\log(n)(1+\varepsilon)$. In order to show that $< S, L > \in ML$, it suffices to show the existence of $\varepsilon > 0$ so that $(m+c_1)(1+\varepsilon) < m + (c_1+c_2)/2$. Since $c_1 < n$ and $m \leq 1.5n$, and $(c_2 - c_1)/2 = n/568$, simple arithmetic shows that by taking $\varepsilon = 1/1420$, the inequality is satisfied.

Suppose $G \notin gap - VC_3$. Then every cover of G is of size $\geq c_2$. According to Theorem 5.11, for every $\varepsilon > 0$ and large enough n, $\log(Pr(S|T_{ML}, \mathbf{p}^*(S,T_{ML}))) < -(m+c_2)\log(n)(1-\varepsilon)$. In order to show that $< S, L > \notin ML$, it suffices to show the existence of $\varepsilon > 0$ so that $(m+c_2)(1-\varepsilon) > m + (c_1+c_2)/2$. Since $c_2 < n$ and $m \leq 1.5n$, and $(c_2 - c_1)/2 = n/568$, simple arithmetic shows that by taking $\varepsilon = 1/1420$ again, the inequality is satisfied. ■

5.4 Other Substitution Models

Our NP hardness result was stated in Neyman's two states model of substitution. What about 4 states DNA, or proteins? It turns out that such extension is not hard. In this section we prove NP-hardness of maximum likelihood reconstruction under the Jukes-Cantor model [15]. This model is a special case of Kimura 2 parameter and 3 parameter models, and of more elaborate models of DNA substitution. The same holds for protein sequences as well.

Suppose we have a c state alphabet, Σ (for DNA sequences, $c = 4$). Let α_e denote a substitution parameter associated with the edge e. In the JC model, there is a certain probability $1 - p_e$ that a character does not change across the edge e. If it does change, the probabilities of

changing to any one of the other $c-1$ characters are equal, $p_e/(c-1)$. The likelihood of S given a tree under this model is defined in a way similar to equation 2:

$$L(S|T,\mathbf{p}) = \sum_{\lambda^n \in \Sigma^{k\times r}} \prod_{e\in E(T)} \frac{p_e}{c-1}^{d_e(\lambda^n)} (1-p_e(\lambda^n))^{k-d_e(\lambda^n)} \tag{3}$$

According to [22], we get for this model relations that are similar to the theorems, lemmata and corollaries in section 4 (with $C^u_{T,pars(S,T)}$ and $C^d_{T,pars(S,T)}$ that are different but have the same order of magnitude). Thus our reduction holds for the JC model, and consequently for all models extending the JC model.

6 Concluding Remarks and Further Research

In this work, we proved that ML reconstruction of phylogenetic trees is computationally intractable. We used the simplest model of substitution – the Neyman two states model [18]. This NP-hardness proof generalizes to the Jukes-Cantor model [15], and then to the Kimura and other models of DNA and protein substitution.

After the extended abstract of this work was submitted and published in RECOMB05, the results in this paper were extended in three directions: We showed that ML remain hard even under the assumption of molecular clock. We proved an initial $1+\varepsilon$ hardness result of approximation for log likelihood (for a rather small ε). We developed an approximation algorithm for log likelihood for special, biologically interesting, sets of inputs. These results were presented in ISMB 2005.

Vertex cover, which is the starting point for our reduction, has a simple 2-approximation algorithm. Maximum parsimony has 2-approximation (and better) algorithms. What about *any constant* approximation algorithms for log likelihood? So far, no constant factor approximations are known. It will be interesting to find a b approximation of log likelihood for some constant $b > 1+\varepsilon$ (for the above ε), or to prove that no such efficient algorithm exists (unless $P = NP$). Finally, it would be nice to identify regions where ML is *tractable*. In this context, we note that it is not even known what is the complexity of *small ML*, where the sequences and the unweighted tree are given, and the goal is to find optimal edge lengths. In practice, local search techniques such as EM or hill climbing seem to perform well, but no proof of performance is known, and multiple maxima [20, 3] shed doubts even on the (worst case) correctness of this approach.

Acknowledgements

We wish to thank Isaac Elias for helpful discussions, and Sagi Snir for reading early drafts of the manuscript.

References

[1] L. Addario-Berry, B. Chor, M. Hallett, J. Lagergren, A. Panconesi, and T. Wareham. Ancestral maximum likelihood of evolutionary trees is hard. *Jour. of Bioinformatics and Comp. Biology*, 2(2):257–271, 2004.

[2] P. Berman and M. Karpinski. On some tighter inapproximability results. *Proc. 26th ICALP*, 1999.

[3] B. Chor, M. D. Hendy, B. R. Holland, and D. Penny. Multiple maxima of likelihood in phylogenetic trees: An analytic approach. *Mol. Biol. Evol.*, 17(10):1529–1541, 2000.

[4] W. Day. The computational complexity of inferring phylogenies from dissimilarity matrix. *Bulletin of Mathematical Biology*, 49(4):461–467, 1987.

[5] W. Day, D. Johnson, and D. Sankoff. The computational complexity of inferring rooted phylogenies by parsimony. *Mathematical Biosciences*, 81:33–42, 1986.

[6] W. Day and D. Sankoff. The computational complexity of inferring phylogenies by compatibility. *Systematic Zoology*, 35(2):224–229, 1986.

[7] I. Dinur and S. Safra. On the importance of being biased (1.36 hardness of approximating vertex-cover). *Annals of Mathematics (accepted)*, 2005.

[8] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[9] J. Felsenstein. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Meth. in. Enzym.*, 266:419–427, 1996.

[10] J. Felsenstein. *Inferring phylogenies*. Sinauer Associates, Sunderland, Massachusetts, 2004.

[11] W. M. Fitch. Toward defining the course of evolution: minimum change for specified tree topology. *Systematic Zoology*, 20:406–416, 1971.

[12] L. Foulds and R. Graham. The steiner problem in phylogeny is np-complete. *Advances in Applied Mathematics*, 3:43–49, 1982.

[13] R.L. Graham and L.R. Foulds. Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computational time. *Math. Biosc.*, 60:133–142, 1982.

[14] J. Hastad. Some optimal inapproximability results. *Journal of ACM*, 48:798–859, 2001.

[15] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. *In H. N. Munro, editor, Mammalian protein metabolism*, pages 21–132, 1969.

[16] M. Karpinski. Approximating bounded degree instances of np-hard problems. *FCT*, 2001.

[17] M. Koshi and R. Goldstein. Probabilistic reconstruction of ancestral nucleotide and amino acid sequences. *Journal of Molecular Evolution*, 42:313–320, 1996.

[18] J. Neyman. Molecular studies of evolution: A source of novel statistical problems. *In S. Gupta and Y. Jackel, editors, Statistical Decision Theory and Related Topics*, pages 1–27., 1971. Academic Press, New York.

[19] M. steel. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of classification*, 9:71–90, 1992.

[20] M. Steel. The maximum likelihood point for a phlogenetic tree is not unique. *Syst. Biol.*, 43:560–564, 1994.

[21] M. Steel and D. Penny. Parsimony, likelihood and the role of models in molecular phylogenetics. *Mol. Biol. Evol.*, 17:839–850, 2000.

[22] C. Tuffley and M. Steel. Link between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology*, 59(3):581–607, 1997.

[23] T. Wareham. On the computational complexity of inferring evolutional trees. *Technical Report 93-01, Department of computer science, Memorial University of Newfoundland*, 1993.

[24] Z. Yang, S. Kumar, and M. Nei. A new method of inferring of ancestral nucleotide and amino acid sequences. *Genetics*, 141:1641–1650, 1995.