Approximability of Selected Phylogenetic Tree Problems

Mathias Hauptmann* Marlis Lamp†

Abstract

We study the approximability of the reconstruction problem of phylogenetic trees with respect to three different cost measures and give the first explicit lower bounds, under standard complexity-theoretic assumptions. For the *Steiner Tree Problem in Phylogeny* (STPP) and the *Generalized Tree Alignment* (GTA) problem, we show that unless $P = NP$, no polynomial-time algorithm can approximate these problems with an approximation ratio below $\frac{359}{358}$. For the *Ancestral Maximum Likelihood* (AML) problem we give a lower bound of $\frac{1577}{1576}$. Furthermore we construct a polynomial-time approximation scheme $(A_\epsilon)_{\epsilon>0}$ for the AML problem, such that for each $\varepsilon > 0$, A_ε is a polynomial time approximation algorithm with ratio $(1+\varepsilon)\cdot(1+\frac{\ln(3)}{2})$. This result is based on the Steiner tree algorithm of Robins and Zelikovsky [RZ00] and on a new exact algorithm for AML instances of constant size. This improves upon recent results by Alon et al. [ACPR08] who gave a 1.78-approximation algorithm for the AML problem.

1 Introduction

Concerning the reconstruction of phylogenetic trees, two major approaches have been considered in the literature: Distance-based methods where only the distances between the n species are given, and character-based methods where for each of the n species the states of m characters are given.

Maximum Likelihood (ML) [F81] and *Maximum Parsimony* (MP) [F71] are

*Dept. of Computer Science, University of Bonn. e-mail:`hauptman@cs.uni-bonn.de`
†Dept. of Computer Science, University of Bonn. e-mail:`lamp@cs.uni-bonn.de`

two well-known optimality criteria that belong to the category of character-based methods. While ML asks for a tree maximizing the likelihood of the given taxa over an arbitrary evolutionary model, MP assumes parsimony as the underlying evolutionary model: The probably that two taxa are closely related is proportional to their similarity.

We consider two versions of MP: The *Steiner Tree Problem in Phylogeny* (STPP) and the problem of *Generalized Tree Alignment* (GTA). STPP is a variant of MP where the underlying genetic distance measure is the *Hamming distance*, that counts the number of differing characters. In GTA, the n species are given as unaligned biological sequences of variable length, so the underlying metric is the *edit distance*. *Ancestral Maximum Likelihood* (AML) is a mixture of MP and ML. AML asks for a tree that maximizes the likelihood of the given species.

STPP, GTA and AML are variants of the *Steiner Tree Problem* where the underlying metric space is some m-dimensional hypercube and the distance measure is the *Hamming distance*, the *edit distance* and the *binary entropy* of the normalized Hamming distance respectively. The *Steiner Tree Problem* asks for a minimum-length tree T connecting a given set S of *terminals* in an underlying metric space (V, d). This is one of the most fundamental network design problems, which is well-known to be NP-hard [K72] and even NP-hard to approximate [CC02]. The currently best known approximation lower bound for the *Steiner Tree Problem* in weighted graphs is 1.01063 [CC02].

In this paper we give the first explicit lower bounds for the approximability of the *Steiner Tree Problem in Phylogeny* (STPP), the *Generalized Tree Alignment* (GTA) and the *Ancestral Maximum Likelihood* (AML). Namely we show that for each $\varepsilon > 0$ it is NP-hard to approximate GTA and STPP with an approximation ratio better than $\frac{359-\varepsilon}{358+\varepsilon}$ and AML with an approximation ratio better than $\frac{1577-\varepsilon}{1576+\varepsilon}$. These results are obtained by constructing approximation-preserving reductions from the *Bounded Degree Vertex Cover Problem* (B-VC) and using explicit lower bounds for the approximability of 5-VC given by Berman and Karpinski [BK98].
Concerning upper bounds, recently Alon et al. [ACPR08] gave a $\frac{16}{9}$-approximation algorithm for the AML problem. This algorithm combines the Steiner Tree approximation algorithm of Berman and Ramaiyer [BR94] with a new

algorithm that efficiently computes optimum Steiner trees for sets of terminals of size at most 4. In this paper we improve upon this result and give a polynomial-time 1.55-approximation algorithm for the AML problem. More precisely, we construct an approximation scheme $(A_\epsilon)_{\epsilon>0}$, such that for each fixed $\epsilon > 0$, A_ϵ is a polynomial-time approximation algorithm for the AML problem with A.R. $(1+\epsilon) \cdot \left(1 + \frac{\ln(3)}{2}\right) \approx (1+\epsilon) \cdot 1.55$. Here our contribution is a family of algorithms $F_k, k \in \mathbb{N}$, such that for each k, F_k is a polynomial-time algorithm that solves to optimality the AML problem for instances with terminal sets up to size k. Plugging this in the algorithm of Robins and Zelikovsky [RZ00] gives the desired algorithm.

The rest of the paper is organized as follows. First we give the precise problem formulations of the STPP, GTA and AML problems. In section 1.2 we refer to previous work. In the sections 2, 3 and 4 we describe our hardness results for the STPP, GTA and AML respectively. The approximation algorithm for the AML problem is described in section 4.3.

1.1 Problem Formulations

In this section, we give some definitions and notations that will be used in the sequel. Furthermore we will give the precise problem formulations of the STPP, GTA and AML.

Let $H^m = \{0,1\}^m$ denote the m-dimensional Boolean hypercube and d_H the *Hamming distance*, i.e. for $x, y \in H^m$ $d_H(x,y) = \sum_{i=1}^m |x_i - y_i|$. Given two strings $x, y \in \{0,1\}^*$, an *alignment* of x and y is a pair of strings $\tilde{x}, \tilde{y} \in \{0,1,\Delta\}^*$ with the following properties:

(i) Deleting all occurrences of Δ from $\tilde{x}$ produces x.

(ii) Deleting all occurrences of Δ from $\tilde{y}$ produces y.

(iii) $\tilde{x}$ and $\tilde{y}$ are of the same length.

A *scoring scheme* is a function $s\colon \{0,1,\Delta\} \times \{0,1,\Delta\} \to \mathbb{R}_+$. The associated *edit distance* d_s is defined as follows: for $x, y \in \{0,1\}^*$, $d_s(x,y) = \min\left\{\sum_{i=1}^{|\tilde{x}|} s(\tilde{x}_i, \tilde{y}_i) \mid \tilde{x}, \tilde{y} \text{ is an alignment of } x \text{ and } y\right\}$.

The notion of an L-reduction was introduced by Papadimitriou and Yannakakis [PY91]. If A and B are optimization problems, then A is *L-reducible* to B with parameters α, β, if there exist two polynomial-time computable functions f, g, such that the following conditions hold: (i) f maps each instance x of A to an instance $f(x)$ of B. (ii) For each instance x of A and solution y to instance $f(x)$ of B, $g(x,y)$ is a solution for instance x of A. (iii) $OPT_B(f(x)) \leq \alpha \cdot OPT_A(x)$. (iv) For each solution y for instance $f(x)$ of B, $|OPT_A(x) - cost(g(x,y))| \leq \beta \cdot |OPT_B(f(x)) - cost(y)|$.

We are now ready to give a precise description of the STPP, GTA and AML problem.

STEINER TREE PROBLEM IN PHYLOGENY (STPP)
Input: A set of n binary sequences $s_1, \dots, s_n$, each of length m
Find: A tree $T = (V, E)$ such that $\{s_1, \dots, s_n\} \subseteq V \subseteq H^m$
Objective: Minimize the *length* $d_H(T) := \sum_{e \in E} d_H(e)$

GENERALIZED TREE ALIGNMENT (GTA)
Input: a set of n binary sequences $s_1, \dots, s_n$, each of length $\leq m$, a scoring scheme $s\colon \{0, 1, \Delta\} \times \{0, 1, \Delta\} \to \mathbb{R}_+$
Find: A tree $T = (V, E)$ such that $\{s_1, \dots, s_n\} \subseteq V \subseteq \{0, 1\}^*$
Objective: Minimize the *mutational length* $d_s(T) := \sum_{e \in E} d_s(e)$.

ANCESTRAL MAXIMUM LIKELIHOOD (AML) Version I
Input: A set of n binary sequences $s_1, \dots, s_n \in H^m$
Find: A tree $T = (V, E)$ such that $\{s_1, \dots, s_n\} \subseteq V \subseteq H^m$
and an assignment of edge probabilities $p\colon E \to [0, 1]$
Objective: Maximize the *overall probability* $\prod_{e \in E} p_e^{d_H(e)} \cdot (1 - p_e)^{m - d_H(e)}$

In [ACH+04] and [ACPR08] it is shown that the AML problem can also be reformulated as a special case of the Steiner Tree Problem in the Boolean hypercube H^m: If we consider the individual edge likelihood $p_e^{d_e}(1 - p_e)^{m - d_e}$ for a given edge e of length $d_e := d_H(e)$ between two taxa with m characters, this term is maximized for $p_e = \frac{d_e}{m}$. Since taking the m-th root and the logarithm are monotone operations that do not change the argument maximizing a function, we are able to reformulate the objective function of AML as a sum and obtain the following discrete variant:

ANCESTRAL MAXIMUM LIKELIHOOD (AML) Version II
Input: A set of n binary sequences $s_1, \ldots, s_n \in H^m$
Find: A tree $T = (V, E)$ such that $\{s_1, \ldots, s_n\} \subseteq V \subseteq H^m$
Objective: Maximize the *overall probability* $p(T)$ of T, where $p(T) := \sum\limits_{e \in E} \frac{d_e}{m} \cdot \log_2(\frac{d_e}{m}) + (1 - \frac{d_e}{m}) \cdot \log_2(1 - \frac{d_e}{m})$

Note that $p(T) = \sum\limits_{e \in E} -H_2(\frac{d_e}{m})$, where $H_2(p) = -p \log_2(p) - (1-p)log_2(1-p)$ is the binary entropy function.

1.2 Previous Work

The NP-hardness of STPP has been shown by Foulds and Graham [FG82], see also [DJS86]. Bern and Plassmann [BP89] proved that already a very restricted version of the *Steiner Tree Problem*, namely the *(1,2)-Steiner Tree Problem*, is APX-hard. This is the *Steiner Tree Problem* restricted to metric instances, where all non-zero distances are 1 or 2. Fernández-Baca and Lagergren showed that the k-restricted STPP is APX-complete for $k \geq 4$ and the k-Steiner ratio for the STPP matches the corresponding ratio for metric spaces defined on networks [FBL98]. Note that for the case $k = 3$ there is a randomized polynomial-time approximation scheme, that solves the 3-restricted *Steiner Tree Problem* with arbitrary precision [PS97]. The GTA problem was shown to be APX-hard by Jiang et al. [JW94]. Their basic idea was to construct a polynomial-time reduction f, that embeds instances I of the STP resulting from Bern and Plassmann's reduction into some hypercube $\{0,1\}^m$ and to show that optimal solutions for these embedded instances $f(I)$ can be assumed not to use any Steiner points from $\{0,1\}^m \backslash f(I)$. The NP-hardness of the AML problem was shown by Addario-Barry et al. [ACH+04]. Alon et al. gave a $\frac{16}{9}$-approximation algorithm for the AML problem [ACPR08].

Both, the previously known hardness results and our new explicit lower bounds for approximablility of the STPP, GTA and AML are essentially based on existing hardness results for the *Bounded Degree Vertex Cover Problem* (B-VC).

B-BOUNDED DEGREE MINIMUM VERTEX COVER (B-VC)
Input: A Graph $G = (V, E)$ of maximum degree $\Delta_G \leq B$
Find: A subset $C \subseteq V$, such that for all $e \in E\; e \cap C \neq \emptyset$
Objective: Minimize $|C|$

The B-VC problem is known to be APX-hard for $B \geq 3$ [PY91]. The best known explicit bounds for non-approximability have been obtained by Berman and Karpinski [BK98].

2 Explicit Lower Bounds for STPP

In 1994, Jiang et al. have already proposed an L-reduction from triangle-free B-VC to STPP [JW94]. Here we analyse this reduction in order to compute an initial approximation lower bound for STPP. For this purpose we will first describe their transformation.

Let $G = (V, E)$ be a triangle-free graph with maximum degree B and $V = \{1, 2, \ldots, m\}$. Without loss of generality, we assume that G is connected. Let $0_{i_1, i_2, \ldots, i_r}$ denote the binary sequence of length m with 0's at the i_j-th positions and 1's at the rest. Then we construct an STPP instance with $S := \{0_{i,j} \mid \{i, j\} \in E\}$. Assuming that G has a vertex cover U of size c, we can construct a phylogeny for S as follows: Connect each sequence $0_{i,j} \in S$ to some 0_k, where $k = i$ or j, and $k \in U$. Afterwards connect the sequences $\{0_i \mid i \in U\}$ to 1^m. Thus each connection has a length of 1, the length of the resulting phylogeny T is $d_H(T) = |E| + c$. We refer the reader to the original paper for more details and verification.

The currently best approximation lower bounds for B-VC are due to Berman, Karpinski [BK98]. Here we need the following theorem.

Theorem 2.1 ([BK98]). *For any $\varepsilon > 0$, it is NP-hard to decide, whether a graph with $140n$ nodes, $12n$ of degree 5 and $128n$ of degree 4, has the minimum size of a vertex cover above $(73 - \varepsilon)n$ or below $(72 + \varepsilon)n$.*

If we combine this result with the reduction from B-VC to *Triangle Free B-VC* that was given by Jiang et al. [JW94], we obtain the following result.

Lemma 2.1. *For any $\varepsilon > 0$, it is NP-hard to decide, whether an instance of Triangle Free 5-VC with $712n$ nodes and $858n$ edges has the minimum size of a vertex cover above $(359 - \varepsilon)n$ or below $(358 + \varepsilon)n$.*

Proof. The reduction given in [JW94] maps each instance $G = (V, E)$ of the B-VC problem to a graph $G' = (V', E')$ with $V' = V \cup \{v_e \mid v \in e \in E\}$ and $E' = \{\{v, v_e\}, \{v_e, w_e\}, \{w_e, w\} \mid e = \{u, v\} \in E\}$, and it is shown there that in polynomial time each vertex cover C' of G' can be transformed into a vertex cover $U' = \{u, w_e \mid u \in U, e = \{u, w\} \in E\}$ such that $U \subseteq V$ is a vertex cover in G and $|U'| \leq |C'|$. If G is a graph with $140n$ nodes, $12n$ of degree 5 and $128n$ of degree 4 then G' consists of $(140 + 2 \cdot 286)n = 712n$ nodes and $3 \cdot \frac{60n+512n}{2} = 858n$ edges, and a vertex cover C of size c in G corresponds to a vertex cover C' of size $c + 286n$ in G'. □

Now we combine the reduction from *Triangle-Free Bounded Degree Vertex Cover* to the STPP given by Jiang et al. with lemma 2.1. This yields the following theorem.

Theorem 2.2. *For any $\varepsilon > 0$, it is NP-hard to decide, whether an instance of STPP with $858n$ taxa has the minimum length of a phylogeny above $(1217 - \varepsilon)n$ or below $(1216 + \varepsilon)n$.*

We increase this initial bound by using a different reduction from B-VC to STPP described below. It turns out that using this reduction, there is no need to require the B-VC instances to be triangle-free anymore.
We start with an arbitrary B-VC instance $G = (V, E)$ with vertex set $V = \{1, 2, \ldots, m\}$. We define $S_G := \{0_{i,j} \mid \{i, j\} \in E\} \cup \{1^m\}$ as the set of taxa of our STPP instance. Note that in contrast to T. Jiang et al.'s reduction, we add the node 1^m to the set of terminals.

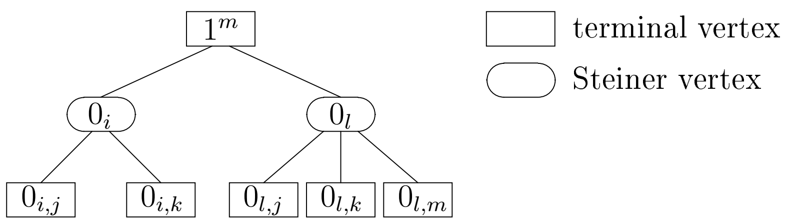

Figure 1: Phylogeny for $S = \{0_{i,j}, 0_{i,k}, 0_{l,j}, 0_{l,k}, 0_{l,m}, 1^m\}$.

Definition 2.1. *Let $G = (V, E)$ be a graph with vertex set $V = \{1, \ldots, m\}$. For each vertex cover $U \subseteq V$ for G we define an associated phylogeny T_U as follows: $L(T_U) := \{0_{i,j} \mid \{i, j\} \in E\}$ is the set of leaves of T_U. Each leaf $0_{i,j}$*

is connected to an inner node 0_k, $k \in \{i,j\} \cap U$ and each inner node 0_k is connected to the root 1^m.

It is easy to see, that T_U is a phylogeny for S_G. See figure 1 for an example. We will now show, that there always exits phylogenies of minimum length that are of the form T_U, where U is a minimum vertex cover for G.

Lemma 2.2. *For each B-VC instance $G = (V, E)$ and each solution T of the corresponding STPP instance $S_G = \{0_{i,j} \mid \{i,j\} \in E\} \cup \{1^m\}$ one can construct a vertex cover U for G in polynomial time, such that $d_H(T_U) \leq d_H(T)$.*

Proof. To show this, we start with an arbitrary phylogeny T for S_G and show that T can be transformed into a phylogeny T_U for S_G without increasing the tree length.
Nodes with sequences that have more than one 0 and are not in S_G are called *bad*. All other nodes are *good*. Without loss of generality we can assume that for each edge e in T $d_H(e) = 1$ and the tree is rooted at 1^m.

Now we will remove all bad nodes in T iteratively from the bottom to

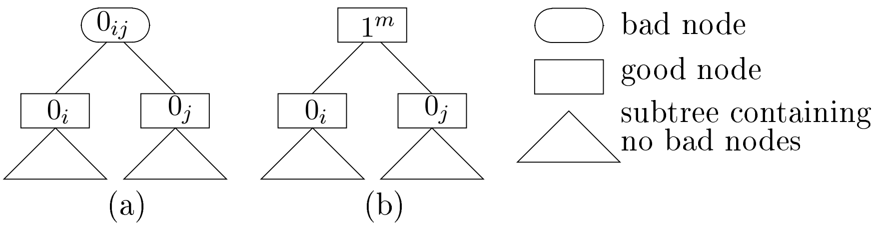

Figure 2: (a) Bad node with two 0's at the lowest level of the tree. *(case 1)* (b) Elimination of the bad node.

the top. Let s be a bad node at the lowest level of the tree. All child nodes of s are good and therefore have at most two 0's. Due to the fact that the Hamming distance of all edges is 1, including the distance between s and each of its children, s cannot have more than three 0's. So it has two or three 0's. We consider these two cases separately.
Case 1: $s = 0_{i,j}$. Since there are no sequences with two 0's and Hamming distance 1 from s, the children of s must have exactly one 0, namely at position i or j. Observe that they have also Hamming distance 1 from 1^m, and so we can connect them to 1^m without increasing the length of the tree.

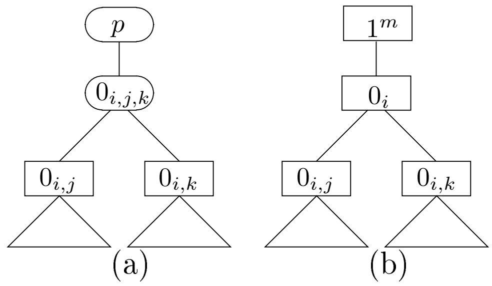

Figure 3: (a) Bad node with three 0's and two children at the lowest level of the tree. *(case 2.1)* (b) Elimination of the bad node.

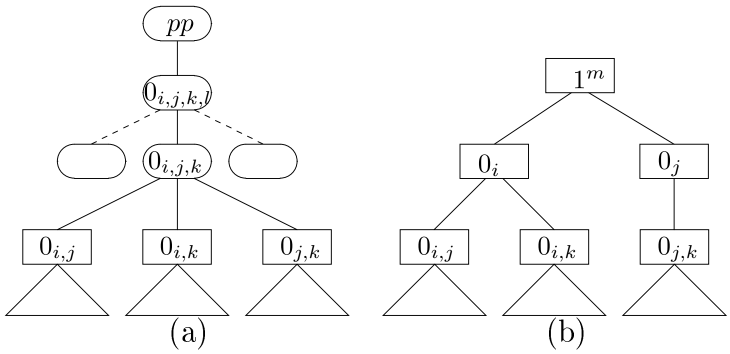

Figure 4: (a) Bad node with three 0's and three children at the lowest level of the tree. *(case 2.2)* (b) Elimination of the bad node.

Now we can delete the bad node s, see figure 2.
Case 2: $s = 0_{i,j,k}$. In this case all children of s must have two 0's, that are at the positions i, j or k. It is easy to see that s can have at most three child nodes. We consider two subcases, depending on the number of children.
Case 2.1: $s = 0_{i,j,k}$ *and* s *has one or two children* . If s has two child nodes they must share one 0-position i. If it has only one we define i as any of the child's 0-positions. Note that all child nodes also have Hamming distance 1 from 0_i. We replace s with 0_i and directly connect this node to 1^m, so that the number of edges is still the same and all edges have Hamming distance 1, see figure 3.
Case 2.2: $s = 0_{i,j,k}$ *and* s *has three children* . In this case, the parent node p of s must have four 0's, because all sequences with two 0's and Hamming distance 1 from s are already spent for the three child nodes of s. This

implies that all siblings of s have three 0's. (Note that they cannot have five 0's, because they are at the same level as s, which means that all nodes below them are good and thus have at most two 0's.) Since all siblings of s share exactly two 0-positions with s, they can have at most two children. By removing all its siblings as described in case 2.1, we achieve that s is the only child of p. Then we connect two children of s, let's say $0_{i,j}$ and $0_{i,k}$, via 0_i to 1^m, which increases the length by 1. Next, we link the remaining child $0_{j,k}$ to 1^m via a 0_j or 0_k, which increases the length by 1 for a second time. Finally we remove s and p and thereby decrease the length by 2, see figure 4 (c). □

Proposition 2.1. *The described mapping is an L-reduction from B-VC to STPP with parameters $\alpha = B + 1$ and $\beta = 1$.*

Proof. Let $G = (V, E)$ be an instance of the B-VC. Let T^* be an optimum solution for the associated instance S_G of the STPP, and let T_U be the phylogeny resulting from application of lemma 2.2 to T^*. Then $d_H(T^*) = d_H(T_U) = |E| + |U|$, which implies $opt(S_G) = opt(G) + |E|$. Since each node in a vertex cover in G can cover at most B edges, $opt(G) \geq \frac{|E|}{B}$ and thus $opt(S_G) = opt(G) \cdot (1 + B)$. This yields $\alpha = B + 1$. Furthermore, if T is an arbitrary solution for the instance S_G of the STPP and T_U is the associated solution resulting from lemma 2.2, then $|d_H(T) - opt(S_G)| \leq |d_H(T_U) - opt(S_G)| = |d_H(T_U) - d_H(T_{U^*})| = |U| - opt(G)$ (where U^* is a minimum vertex cover in G) and thus $\beta = 1$. □

To finish our reasoning, it remains to perform the accounting. Let G be the 5-VC instance constructed by Berman and Karpinski [BK98]. It has $12n$ degree-5 nodes and $128n$ degree-4 nodes and the difficult question is, whether G has the minimum size of a vertex cover U above $(73 - \varepsilon)n$ or below $(72 + \varepsilon)n$. Each of the $286n$ terminals corresponding to edges in G is connected to some 0_i, $i \in U$, that is connected to 1^m. Since each connection costs 1, the cost of the resulting phylogeny is at least $(73 + 286 - \varepsilon)n$, or at most $(72 + 286 + \varepsilon)n$.

Theorem 2.3. *For any $\varepsilon > 0$, it is NP-hard to decide whether an instance of STPP with $286n$ taxa has the minimum length of a phylogeny above $(359-\varepsilon)n$ or below $(358 + \varepsilon)n$.*

3 Generalized Tree Alignment

The reduction from B-VC to STPP given by Jiang et al. [JW94], which we described in the preceding section, also gives a reduction from B-VC to GTA, if we use the score scheme presented in table 1. Thus the approximation lower bound of 1.0027 which is obtained by combining this reduction with the hardness results of Berman and Karpinski for 5-VC [BK98] as shown in section 2 also holds for GTA. Here we show, that the L-reduction from B-VC to STPP constructed in section 2 also works for the GTA problem. This gives the new lower bound of 1.0028 also for GTA.

	0	**1**	Δ
0	0	1	2
1	1	0	2
Δ	2	2	0

Table 1: score scheme

For the sake of completeness, we will first describe how to transform an arbitrary *Bounded Degree Vertex Cover Problem* into a special GTA instance. Let $G = (V, E)$ be a graph with degree bounded by B. We number the vertices of V consecutively from 1 to m. As input sequences of the GTA instance, we choose for each edge $\{i, j\}$ in G a '1'-sequence of length m, with only two zeros at the positions i and j. Finally we add 1^m to the set S_G of input sequences.

Let $U \subset V$ be a minimum vertex cover for G. We build a phylogeny T of minimum (mutational) length for S_G as described in the previous section. We define the score scheme s as in table 1. We still have to prove, that the sum of distances along the edges of T is minimum. To show this, we start with an arbitrary phylogeny T' for S_G and show, that T' can be transformed into T without increasing the tree length.

Again we divide the nodes in T' into *bad* nodes, that are neither in S_G nor of form 0_i and *good* nodes, that are not bad. The following lemma allows to restrict our considerations to trees T' with additional structural properties.

Lemma 3.1. *There is an polynomial-time algorithm that transforms a given phylogeny T into a phylogeny T' of length $d_s(T') \leq d_s(T)$, such that the following properties hold.*

(i) *For each edge e in T', $d_s(e) = 1$.*

(ii) *All node sequences in T' have length m.*

(iii) *Each bad node in T' has at least two children.*

(iv) *Each sequence appears at most once in T'.*

Proof. Property (i) can be achieved by deleting each edge that is longer than 1. This separates T' into two components. One of them contains the input sequence 1^m and the other one must contain any input sequence of form $0_{i,j}$. If not, the other component would not contain any input sequence and could be removed. We reconnect the two components by linking $0_{i,j}$ and 1^m to some sequence 0_k, $k \in U \cap \{i, j\}$. Since both new edges have length 1 and the length of the original edge was at least 2, this does not increase the length of the tree. Since the score of a gap is 2, (ii) follows from (i). Bad nodes with only one child can be deleted with both connecting edges of length 1 and the two disconnected components can be reconnected as before with at most two new edges, each of length 1. Thus afterwards (iii) holds. (iv) can easily be achieved by moving edges and removing the isolated duplicates. □

Now we can remove all bad nodes in T' iteratively from the bottom to the top, assuming that the tree is rooted at 1^m. We refer the reader to the proof of proposition 2.1 in section 2, because due to lemma 3.1 all distances occurring in T' and T equal the Hamming distance.

Proposition 3.1. *The described mapping is an L-reduction from B-VC to GTA with parameters $\alpha = B + 1$ and $\beta = 1$.*

Combining this result with the 5-VC approximation lower bound of Berman and Karpinski [BK98] we obtain the following theorem.

Theorem 3.1. *For any $\varepsilon > 0$, it is NP-hard to decide, whether an instance of GTA with $286n$ input sequences has the minimum mutational length of a phylogeny above $(359 - \varepsilon)n$ or below $(358 + \varepsilon)n$.*

4 Ancestral Maximum Likelihood

In the general version of AML given above, we have to optimize over tree topologies, sequence assignments and edge probabilities. In 2000, Pupko et al. developed a dynamic programming solution for a special version of AML, where the topology and the edge lengths are given as part of the input [PPSG04]. Three years later Addario-Barry et al. proved that its general version is NP-hard [ACH+04], by using a reduction from *Vertex Cover*. Here we use these results to show that AML is even APX-hard and to compute an explicit approximation lower bound.

4.1 APX-Hardness

For proving the APX-hardness of AML we use the existing polynomial reduction given by Addario-Barry et al. from Vertex Cover (VC) to AML. In their construction, in a first step the *Vertex Cover Problem* is reduced to a special case, the *Vertex Cover Problem* in h-bajan graphs. An h-bajan graph of G basically consists of h disjoint copies of G connected to each other along the edges of G. This intermediate step was necessary in order to obtain a polynomial-time reduction to the AML, since in the case of the binary entropy cost function used in AML the construction, described in the proof of lemma 3.1 does not apply to trees resulting from general *Vertex Cover* instances (cf. [ACH+04]). Here we will show that *Vertex Cover* on h-bajan graphs remains APX-hard. Subsequently one can L-reduce from *Vertex Cover* on h-bajan graphs to AML.

We begin with the definition of h-bajan graphs. For any integer $h > 1$, the h-bajan graph $B(G, h)$ of a graph G consists of h isomorphic copies of G. The copy of vertex u in the ith copy of G is denoted as u^i. Two vertices u^i and u^j are connected in $B(G, h)$, if u and v are connected in the original graph G.

Definition 4.1 (h-bajan [ACH+04]). *Let $G = (V, E)$ be a graph and $h > 1$ be an integer. The h-bajan graph of G is defined as follows: $B(G, h) := (V_B, E_B)$, where $V_B := \bigcup_{i=1}^{h} \{v^i \mid v \in V\}$ and $E_B := \{\{u^i, v^j\} \mid 1 \leq i, j \leq h \wedge \{u, v\} \in E\}$.*

Addario-Barry et al. have shown, that Vertex Cover on h-bajan graphs remains NP-hard. Here we need the following lemma:

Lemma 4.1. *For any $h > 1$, Vertex Cover on h-bajan graphs is APX-hard.*

Proof. Let $G = (V, E)$ be an arbitrary graph, $h > 1$ and $B(G, h) = (V_B, E_B)$ be the h-bajan graph of G. Obviously B can be created in polynomial time. Suppose that U is a vertex cover in B. Since the edges inside each copy of G can be covered only by vertices inside this copy, U can be written as $\bigcup_{i=1}^{h} U_i$, where each U_i contains only vertices of the i-th copy of G. We can transform each vertex cover of B into a so called *normalized* vertex cover without increasing its size by choosing the minimal $|U_i|$ and carrying it to the other copies of G. Let U_G be a vertex cover of G of size c. The size of a corresponding vertex cover in B is $h \cdot c$, so it depends linearly on c. $\square$

Let $G = (V, E)$ be an h-bajan graph with $V = \{1, \ldots, m\}$ and $|E| = n$. One can construct an instance S of AML, such that $|S| = |E| + 1 = n + 1$ and $S \subseteq \{0, 1\}^m$. Again let $0_{i_1,i_2,\ldots,i_r}$ denote the binary sequence of length m with 0's at the i_j-th positions and 1's at the rest. Then we define S as $\{0_{i,j} \mid \{i, j\} \in E\} \cup \{1^m\}$. Addario-Barry et al. have shown, that any phylogenetic tree T for S can be transformed into a tree T' of the following form, without increasing its likelihood:

- For any edge e in T': $d_e = 1$.
- All nodes connected to 1^m are internal vertices of the form 0_i.
- All leaves of T' are in $S \setminus \{1^m\}$.

In the sequel we consider 1^m as the root of T'. The set $\{i \mid 0_i \in V(T)\}$ corresponds to a vertex cover in G.
Let U be a minimal vertex cover in G and $c = |U|$. The likelihood of the corresponding phylogeny can be computed as follows: There are n edges from the leaves to the inner nodes and c edges from the inner nodes to the root. For each edge e in T, $d_e = 1$. So the likelihood of the phylogeny is $(n + c) \cdot (-H(\frac{1}{m}))$, which is linear in c. Theorem 4.1 follows.

Theorem 4.1. *AML is APX-hard.*

4.2 Explicit Lower Bounds

In this subsection we compute an explicit lower bound for AML. Therefore we use a combination of the above described reduction from vertex cover

to AML and the lower bounds for the approximability of 5-VC given by P. Berman and M. Karpinski [BK98].
Let $G = (V, E)$ be a graph with $12n$ degree-5 nodes and $128n$ degree-4 nodes. For G it is NP-hard to decide, if the minimum size of a vertex cover U is above $(73 - \varepsilon)n$ or below $(72 + \varepsilon)n$.
Therefore, the 2-bajan graph $B(G, 2)$ of G has $2{\cdot}n$ nodes and $(3{\cdot}2-1){\cdot}286n = 1430$ edges and the difficult question is, whether it has a vertex cover of above $(73 - \varepsilon)2n$ or below $(72 + \varepsilon)2n$ vertices.
Let S be the corresponding AML instance with $1431n$ terminals. It is NP-hard to decide, if the maximum likelihood of a phylogeny for S is above $(1577-\varepsilon)(-H(\frac{1}{280n}))$ or below $(1576+\varepsilon)(-H(\frac{1}{280n}))$. Thus the approximation lower bound for AML is $\frac{1576+\varepsilon}{1577+\varepsilon} \approx 1.00063$.

Theorem 4.2. *For any $\varepsilon > 0$, it is NP-hard to decide, whether an instance of AML with $1431n$ taxa has the maximum likelihood of a phylogeny above $(1577 - \varepsilon)(-H(\frac{1}{280n}))$ or below $(1576 + \varepsilon)(-H(\frac{1}{280n}))$.*

4.3 Approximate Solutions

In this subsection we describe our new approximation scheme $(A_\varepsilon)_{\varepsilon>0}$ for the AML problem. For each $\varepsilon > 0$, A_ε will be a polynomial-time approximation algorithm with ratio $(1+\varepsilon)\cdot\left(1 + \frac{\ln(3)}{2}\right)$. Recall that the Steiner tree approximation algorithm k-LCS of [RZ00] consists of a greedy algorithm that starts with a minimum spanning tree T_0 and iteratively inserts optimum Steiner trees T_i for subsets of the terminal set of size at most k. Thus, in order to get the same results for the AML problem it is sufficient to construct a family of polynomial-time algorithms $F_k, k \in \mathbb{N}$ such that for each k, F_k solves the AML problem for terminal sets up to size k.

We will now describe how to construct such a family $F_k, k \in \mathbb{N}$. First recall that the probability of an edge e for an instance $S \subseteq H^m$ of the AML problem is $-H_2\left(\frac{d_e}{m}\right)$. We observe that edge probabilities can take only $m+1$ different values $-H_2(\frac{d_e}{m}) \in \left\{0, -H_2(\frac{1}{m}), \ldots, -H_2(1)\right\}$.

Pupko et al. [PPSG04] gave a polynomial-time algorithm $\mathcal{A}_{glt}$ which constructs a phylogeny T for a given set of leaves $S \subseteq H^m$, a given topology $\mathcal{T}$ and given edge lengths $d(e)$. More formally, a *topology* $(\mathcal{T}, l)$ *for a given set of leaves* consists of a tree $\mathcal{T} = (V, E)$ and a bijection $l\colon L(\mathcal{T}) \to S$ between the *set of leaves* $L(\mathcal{T})$ of $\mathcal{T}$ and the set S.

A *topology with given edge lengths* for a given set of leaves S is a topology $(\mathcal{T}, l)$ for S with a labeling $d\colon E \to [0, \infty)$, where $\mathcal{T} = (V, E)$. Algorithm $\mathcal{A}_{glt}$ on input $S, \mathcal{T}, l, d$ either constructs an assignment $I\colon V \setminus L(\mathcal{T}) \to H^m$ of points from H^m to the inner nodes of $\mathcal{T}$ such that this assignment is consistent with the edge lengths and the labeling l of the leaves - or $\mathcal{A}_{glt}$ retuns "no" in case no such assignment exists. Here we call such an assignment I *consistent* if for all $e = \{u, v\} \in E$ with $u, v \in V \setminus L(\mathcal{T})$, $d_H(I(u), I(v)) = d(e)$ and for all $e = \{u, v\} \in E$ with $u \in L(\mathcal{T})$ and $v \in V \setminus L(\mathcal{T})$, $d(e) = d_H(I(v), L(u))$.

Let $T_{glt} = T_{glt}(S, \mathcal{T}, l, d)$ denote the resulting tree, i.e. $T_{glt} = (V_{glt}, E_{glt})$ is a tree, such that $S \subseteq V_{glt} \subseteq H^m$.

Algorithm F_k gets as an input a set of terminals $S \subseteq H^m$ of size $|S| \leq k$. F_k simply enumerates all topologies $(\mathcal{T}, l)$ for the given set of leaves S and for each topology all assignments of lengths from $\{0, 1, \ldots, m\}$ to the edges of $\mathcal{T}$. For each triple $\mathcal{T}, l, d$ the algorithm F_k uses Pupko et al.'s algorithm $\mathcal{A}_{glt}$ to compute a tree $T_{glt}(\mathcal{T}, l, d)$. Finally F_k returns one of these trees maximizing the overall probability $p(T_{glt})$.

Algorithm $\mathbf{F_k}$

Input: set of terminals $S \subseteq H^m$ of size $|S| \leq k$
Output: optimum Steiner tree T for S in H^m

(1) **for all** topologies with edge lengths $\mathcal{T}, l, d$ for S:
(2) $\quad T_{\mathcal{T},l,d} := \mathcal{A}_{glt}(\mathcal{T}, l, d)$;
(3) $T := \arg\max_{\mathcal{T},l,d} p(T_{\mathcal{T},l,d})$;

(4) **return** T;

Figure 5: Algorithm $\mathbf{F_k}$

For a given set S of cardinality $|S| \subseteq k$, we estimate the number t_S of topologies for the set of leaves S as follows. A tree with at most k leaves has at most $k-1$ inner nodes and at most $2k-2$ edges, thus $t_S = k^{O(k)}$. For each such topology $\mathcal{T}$, the number of assignments of lengths to edges can be bounded by $m^{O(k)}$ which is a polynomial bound in m if k is constant. Thus we obtain the following result.

Theorem 4.3. *For terminal sets $S \subseteq H^m$ of size at most k, algorithm F_k solves the AML problem to optimality. Its running time is $T_k(m) = m^{O(k)}$*

which is polynomial in m if k is constant.

Corollary 4.1. *There is a polynomial-time approximation scheme $(A_\varepsilon)_{\varepsilon>0}$ for the AML problem with approximation ratio $(1+\varepsilon)\cdot\frac{\ln(3)}{2}\approx(1+\varepsilon)\cdot 1.55$.*

5 Conclusion

In this paper we have given the first explicit lower bounds for the approximability of the STPP, GTA and AML problems. Furthermore, based on the Robins-Zelikovsky algorithm for the *Steiner Tree problem* we construct a $\left(1+\frac{\ln(n)}{2}\right)$-approximation scheme for the AML problem.

Several open questions remain. We believe that the lower bounds presented here can be improved by using more direct reductions from MAX E3-LIN2 in the style of Chlebík, Chlebíková [CC02]. Note that their reduction produces edge-weighted instances of the *Steiner Tree Problem* which cannot be directly embedded into sufficiently small hypercubes. It would also be interesting to extend the hardness result to the case of arbitrary alphabets of constant size.

References

[ACH+04] 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, Vol. 2, No. 2, pp. 257–271, 2004.

[ACPR08] N. Alon, B. Chor, F. Pardi, and A. Rapoport. Approximate maximum parsimony and ancestral maximum likelihood. Technical report, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2008.

[BK98] P. Berman and M. Karpinski. On some tighter inapproximability results, further improvements. *Electronic Colloquium on Computational Complexity (ECCC)*, 5(65), 1998.

[BK03] P. Berman and M. Karpinski. Improved approximation lower bounds on small occurrence optimization. *Electronic Colloquium on Computational Complexity (ECCC)*, 10(8), 2003.

[BR94] P. Berman and V. Ramaiyer. Improved approximations for the steiner tree problem. *Journal of Algorithms*, 17(3):381–408, 1994.

[BP89] M. Bern and P. Plassmann. The steiner problem with edge lengths 1 and 2. *Inf. Process. Lett.*, 32(4):171–176, 1989.

[CC02] M. Chlebík and J. Chlebíková. Approximation hardness of the steiner tree problem on graphs. In *SWAT '02: Proceedings of the 8th Scandinavian Workshop on Algorithm Theory*, pages 170–179, London, UK, 2002. Springer-Verlag.

[DJS86] W.H.E. Day, D.S. Johnson and D. Sankoff. The computational complexity of inferring rooted phylogenies. *Math. Biosci.* 81, pp. 33–42, 1986

[F81] J. Felsenstein. *Evolutionary trees from DNA sequences: a maximum likelihood approach.* Journal of Molecular Evolution, 17(6):368-76, 1981.

[FBL96] D. Fernández-Baca and J. Lagergren. A polynomial-time algorithm for near-perfect phylogeny. In *ICALP '96: Proceedings of the 23rd International Colloquium on Automata, Languages and Programming*, pages 670–680, Berlin /Heidelberg, Germany, 1996. Springer-Verlag.

[FBL98] D. Fernández-Baca and J. Lagergren. On the approximability of the steiner tree problem in phylogeny. *Discrete Applied Mathematics*, 88(1-3):129–145, 1998.

[F71] W. Fitch. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416, 1971.

[FG82] L. Foulds and R. Graham. The steiner tree problem in phylogeny is np-complete. *Advances in Applied Mathematics*, 3:43–49, 1982.

[JW94] T. Jiang and L. Wang. On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1(4):337–348, 1994.

[K72] R. Karp. Reducibility among combinatorial problems. In R. Miller and J. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

[PY91] C. Papadimitriou and M. Yannakakis. Optimization, approximation, and complexity classes. *Journal of Computer System Sciences*, 43(3):425–440, 1991.

[PPSG04] I. Pe'er, T. Pupko, R. Shamir and R. Sharan. Incomplete Directed Perfect Phylogeny. SIAM J. Comput. 33(3): 590–607, 2004

[PS97] J. Prömel and A. Steger. RNC-approximation algorithms for the steiner problem. In *STACS '97: Proceedings of the 14th Annual Symposium on Theoretical Aspects of Computer Science*, pages 559–570, London, UK, 1997. Springer-Verlag.

[RZ00] G. Robins and A. Zelikovsky. Improved steiner tree approximation in graphs. In *SODA 2000: Proceedings of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 770–779, 2000.