HAL open science

When Three Trees Go to War

Leo van Iersel, Mark Jones, Mathias Weller

▶ To cite this version:

Leo van Iersel, Mark Jones, Mathias Weller. When Three Trees Go to War. 2023. hal-04013152v3

HAL Id: hal-04013152
https://hal.science/hal-04013152v3

Preprint submitted on 19 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

When Three Trees Go to War

Leo van Iersel[1], Mark Jones[1], and Mathias Weller[2,3]

[1] Delft Institute of Applied Mathematics, Delft University of Technology – The Netherlands
[2] Technische Universität Berlin – Germany
[3] CNRS, LIGM (UMR 8049), Université Gustave Eiffel – France

Correspondence: mathias.weller@cnrs.fr

Abstract

How many reticulations are needed, in the worst case, for a rooted phylogenetic network to display a given set of k rooted binary phylogenetic trees on n leaves? For $k = 2$, Baroni, Semple, and Steel [Ann. Comb. 8, 391–408 (2005)] showed that the answer is $n - 2$. Here, we show that, for $k \geq 3$ the answer is at least $(3/2 - \epsilon)n$. Concretely, we prove that, for each $\epsilon > 0$, there is some $n \in \mathbb{N}$ such that three n-leaf caterpillar trees can be constructed in such a way that any network displaying these caterpillars contains at least $(3/2 - \epsilon)n$ reticulations. The case of three trees is interesting since it is the easiest case that cannot be equivalently formulated in terms of agreement forests. Instead, we base the result on a surprising lower bound for multilabelled trees (MUL-trees) displaying the caterpillars. Indeed, we show that one cannot do (more than an ϵ) better than the trivial MUL-tree resulting from a simple concatenation of the given caterpillars. The results are relevant for the development of methods for the Hybridization Number problem on more than two trees. This fundamental problem asks to construct a rooted phylogenetic network with a minimum number of reticulations displaying a given set of rooted phylogenetic trees.

Keywords: Phylogenetic Networks, Tree Containment, Caterpillar Tree, Hybridization Number

1 Introduction

A fundamental task of evolutionary analysis is to construct a phylogeny for a set of taxa depicting their ancestral relations. While many biological studies are content with the simplification that this phylogeny is a tree, there are circumstances, such as the presence of horizontal gene transfer (observed in many bacteria [DARM08]) or hybridization (common among plant species and also observed among animals [Mal05]), that require constructing phylogenetic *networks* which, in contrast to trees, allow modeling such "reticulate" evolution. Mathematically, a rooted phylogenetic network is a directed acyclic graph (DAG) with a single root and leaves bijectively labelled by the elements of a set X, modeling the considered taxa. In this paper, we will only consider phylogenetic networks that are rooted and binary. In such networks, the indegree-2 nodes represent reticulate evolutionary events and are called *reticulations*.

The construction of the most parsimonious (that is, containing the least amount of reticulations) network that is still "compatible" with a given set of trees is modeled by the Hybridization Number problem, which is well understood for the special case of having exactly two input trees [BS07; BSS05; Bar+05; IK11; Kel+12]. To formalize this problem, we say that a network *displays* a tree if this tree can be obtained from a subgraph of the network by suppressing nodes with exactly one incoming and exactly one outgoing arc. The Hybridization Number problem then asks for a smallest network (in terms of the number of reticulations) displaying all input trees. Throughout this paper, we will focus on the simplified version of this problem where all input trees and the output network are required to be binary and all have the same set of leaf labels.

Unfortunately, many observations made for this case do not generalize to more than two input trees. One such observation is that for any two trees with n leaves, there is always a network with $n-2$ reticulations displaying the two trees. This bound is tight because two "inverse" caterpillar trees need exactly $n-2$ reticulations [BSS05]. In this work, we show that, for three or more trees, at least $(3/2-\epsilon)n$ reticulations may be required, even if the trees are caterpillars. See Fig. 1 for an example of our construction. This result represents a first lower bound for more than two trees that improves upon the $n-2$ bound. In particular, it refutes the tempting conjecture that n reticulations are sufficient to display any set of three phylogenetic trees. If the bound of n had held, it would have had positive consequences for the development of methods for the Hybridization Number problem, by bounding the worst-case complexity of subnetworks that need to be considered inside an algorithm.

To prove the $(3/2-\epsilon)n$ bound, we fist derive a corresponding bound for "multilabelled trees" (MUL-trees), that is, trees in which each leaf has one label, but each label may be used more than once. Again, the goal is to find a smallest (in terms of the number of leaves) MUL-tree displaying all input trees. Surprisingly, we show that, given at most three caterpillars, one cannot do better (up to an ϵ) than the trivial MUL-tree that simply concatenates the given caterpillars. More precisely, we show that, for each $\epsilon > 0$ and $t \leq 3$, there is some $n \in \mathbb{N}$, and t caterpillars with n leaves, such that any MUL-tree displaying the caterpillars has at least $(t-\epsilon)n$ leaves. This is very close to the upper bound of $t \cdot n$ for any set of t trees, which holds because the t trees can simply be concatenated into a single MUL-caterpillar with $t \cdot n$ leaves.

Upper bounds on the number of reticulations needed to display a set of trees follow from results on "universal tree-based" networks [BS18], which are, roughly speaking, networks that can be obtained from *any tree* on the same set of leaves by subdividing arcs of the tree and adding arcs between subdividing nodes [FS15; Hay16; Zha16]. Bordewich and Semple [BS18] showed that such a network has $\Theta(n \log n)$ reticulations. The lower bound does not (directly) carry over to our question, but the upper bound does. Concretely, Bordewich and Semple [BS18] proved that any network displaying all phylogenetic trees on n leaves needs at least $O(n \log n)$ reticulations.

The structure of this paper is as follows. After the preliminaries in Section 2, we prove the bound for MUL-trees in Section 3 and the bound for networks in Section 4, concluding with some open problems in Section 5.

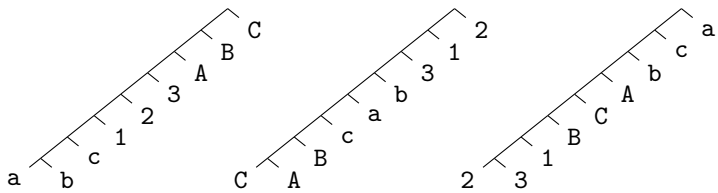

Figure 1. Three caterpillar trees with 9 leaves, resulting from Construction 1 for $n = 3^i = 9$. Lemma 2 implies that any MUL-tree displaying these caterpillars needs at least 19 leaves. Lemma 3 further implies that any network displaying these caterpillars needs at least 2 reticulations. While this second number does not seem particularly surprising, the strength of Lemma 3 is in the *asymptotic bound* it provides for growing n.

2 Preliminaries

We will deal with sequences of letters over an unspecified alphabet. To differentiate such letters from variable names (even those referring to letters), we will typeset them in typewriter font such as a, 3, B. For all sequences s' that can be produced from s by removing zero or more letters, we say that s' is a *subsequence* of s and we write $s' \trianglelefteq s$. We use $\circ$ to denote the usual concatenation operator on sequences, where $s \circ s'$ denotes the result of writing out s' after s. For letters a and b, we write a $\leq_s$ b if the last occurence of a preceeds the first occurence of b in s, that is, some prefix of s contains all occurences of a but no occurence of b.

We also deal with (*binary*, *phylogenetic*) *MUL-networks*, which are directed, acyclic graphs (DAGs) with only the following types of nodes: (1) a unique source (called the *root*) with out-degree zero or two; (2) sinks (called *leaves*) with in-degree at most one, labeled using a function $\mathcal{L}$ from the set of leaves to some set of labels; (3) in-degree one and out-degree two nodes (*tree nodes*); and (4) in-degree two, out-degree one nodes (*reticulation nodes* or *reticulations*). A *MUL-tree* is a MUL-network without reticulations. A *network* is a MUL-network whose labelling function $\mathcal{L}$ is injective. A *tree* is a network without reticulations. A *caterpillar tree* (or just a *caterpillar*) is a tree in which each node is either a leaf or the parent of a leaf. If X is a MUL-network, a subset of nodes in a MUL-network, or a sequence, then $\mathcal{L}(X)$ is the *set* (not the multiset) of labels/letters occurring in X. If X is a MUL-network, then $n(X)$ denotes the number of leaves in X. If X is a sequence, then $n(X)$ denotes its length.

We use the following correspondence between sequences and caterpillars. We say that sequence s *corresponds* to caterpillar P if the elements of s are exactly the leaf labels of P and these labels are ordered in s by decreasing distance from the root in P. Observe that each caterpillar has exactly two corresponding sequences since it has exactly two leaves with the same distance from the root. Conversely, each sequence has exactly one corresponding caterpillar. See Fig. 1 for an example of three caterpillars with corresponding sequences `abc123ABC`, `CABcab312` and `231BCAbca`.

Let N be a MUL-network and let u be a node of N. If u has a descendant v in N (that is, N contains a directed u-v-path), then we write $v \leq_N u$ and we call u an *ancestor* of v. Note that the "$\leq_X$"-relations for MUL-networks and sequences X naturally extend to sets of nodes/letters, that is, $P <_X Q$ if $p <_X q$ for all $p \in P$ and $q \in Q$. For a set L of nodes of N, the MUL-network of N *induced* by L (written $N[L]$) is the result of removing all nodes u of N that have no descendant in L followed by the exhaustive contraction of indegree-one outdegree-one nodes onto their respective parents (this last operation is also called "supression"). If L is a set of labels, then $N[L]$ is the MUL-network of N induced by all nodes with a label in L. The result of removing all nodes x from N with $x \not\leq_N u$ is denoted by N_u and, if u is a reticulation and N_u does not contain any reticulations of N, then u is called a *lowest reticulation*. If N is a MUL-tree and x and y are nodes in N, then the *lowest common ancestor* (LCA) of x and y in N is the unique minimum with respect to "$\leq_N$" of all nodes u of N such that N_u contains both x and y.

An *embedding* of a MUL-network T into a MUL-network N is a function ϕ that maps the nodes of T to nodes of N and the arcs of T to directed paths in N such that

1. the paths in the image of ϕ are arc-disjoint;
2. for each arc uv of T, $\phi(uv)$ starts in $\phi(u)$ and ends in $\phi(v)$ in N.

i	0	1	2	3
X_i	λ	a1A	abc123ABC	abcdefghi123456789ABCDEFGHI
Y_i	λ	Aa1	CABcab312	IGHCABFDEighcabfde978312645
Z_i	λ	1Aa	231BCAbca	564897231EFDHIGBCAefdhigbca

Figure 2. Example of Construction 1. For $i = 1$, the functions r_1, r_2, and r_3 map λ to a, 1, and A, respectively. For $i = 2$, they map $\{\mathtt{a}, 1, \mathtt{A}\}$ to $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$, $\{1, 2, 3\}$, and $\{\mathtt{A}, \mathtt{B}, \mathtt{C}\}$, respectively. In particular, sequence X_2 is given by $r_1(\mathtt{a1A}) \circ r_2(\mathtt{a1A}) \circ r_3(\mathtt{a1A}) = \mathtt{abc123ABC}$, while Y_2 is given by $r_3(\mathtt{Aa1}) \circ r_1(\mathtt{Aa1}) \circ r_2(\mathtt{Aa1}) = \mathtt{CABcab312}$. For $i = 3$, r_1 maps $\{\mathtt{a}, \mathtt{b}, \mathtt{c}, 1, 2, 3, \mathtt{A}, \mathtt{B}, \mathtt{C}\}$ to $\{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e}, \mathtt{f}, \mathtt{g}, \mathtt{h}, \mathtt{i}\}$, and analogously for r_2, r_3.

We say that MUL-network N *displays* MUL-network T if there is an embedding of T into N.

The *backbone* of a caterpillar is the path containing all edges not incident to a leaf. The *backbone* of an embedding ϕ of a caterpillar P in a MUL-network N is the path obtained by merging the paths $\phi(e)$ for all edges e on the backbone of P.

3 Lower Bound on MUL-Trees

In this section, we construct a family $\mathcal{C}$ of triples of caterpillars such that for any $\epsilon > 0$, the familiy $\mathcal{C}$ contains a triple (C_1, C_2, C_3) of n-leaf caterpillars (where n depends on the choice of ϵ) such that any MUL-tree displaying all three caterpillars has at least $(3 - \epsilon)n$ leaves. As a byproduct, we show the existance of a family of *pairs* of caterpillars with a $(2 - \epsilon)n$ lower-bound on the leaf-number in any displaying MUL-tree.

A *relabeling* is a function mapping a label to another label and we allow applying relabelings to sets of labels, sequences and (MUL-)trees in the natural way.

Construction 1. *Let $\mathcal{C}_0 = (X_0, Y_0, Z_0)$ denote the triple of sequences on a single label λ. For each $i > 0$, we recursively construct a triple $\mathcal{C}_i = (X_i, Y_i, Z_i)$ of sequences of length 3^i as follows: Let r_1, r_2, and r_3 be relabelings defined on the labels of $\mathcal{C}_{i-1}$ with disjoint images. Then,*

1. $X_i := r_1(X_{i-1}) \circ r_2(X_{i-1}) \circ r_3(X_{i-1})$
2. $Y_i := r_3(Y_{i-1}) \circ r_1(Y_{i-1}) \circ r_2(Y_{i-1})$
3. $Z_i := r_2(Z_{i-1}) \circ r_3(Z_{i-1}) \circ r_1(Z_{i-1})$

Note that $\mathcal{L}(r_1(X_{i-1}))$, $\mathcal{L}(r_1(Y_{i-1}))$ and $\mathcal{L}(r_1(Z_{i-1}))$ are identical and we refer to this set by A_i. Similarly, we abbreviate $B_i := \mathcal{L}(r_2(X_{i-1}))$ and $C_i := \mathcal{L}(r_3(X_{i-1}))$.

It turns out that sequences constructed by Construction 1 have very short common subsequences.

Proposition 1. *Let $i > 0$ and let (X_i, Y_i, Z_i) be a triple of sequences constructed by Construction 1. Let $k \in \{1, 2, 3\}$ and let s_k be a common subsequence of any k of the three sequences. Then, $|s_k| \leq (4 - k)^i$.*

Proof. Clearly, the claim trivially holds for $k = 1$ so we consider $k \in \{2, 3\}$ in the following.

Case 1: $k = 3$. The proof is by induction on i. For $i = 0$, all three sequences contain a single label λ so the claim is trivially true. Suppose in the following that the claim holds for $i-1$. Let s_3 be a common subsequence of X_i, Y_i, and Z_i. Since all labels in the image of r_1 preceed all labels in the image of r_3 in X_i and all labels in the image of r_3 preceed all labels in the image of r_1 in Y_i, we know that s_3 does not contain labels of the images of both r_1 and r_3. Similarly, it can be seen that s_3 cannot contain labels of any two of r_1, r_2, and r_3. Thus, without loss of generality, s_3 consists only of labels of r_1, implying that s_3 is a common subsequence of the result of removing all labels of r_2 and r_3 from X_i, Y_i, and Z_i, that is, s_3 is a common subsequence of $r_1(X_{i-1})$, $r_1(Y_{i-1})$ and $r_1(Z_{i-1})$. But then, $r_1^{-1}(s_3)$ is a common subsequence of X_{i-1}, Y_{i-1} and Z_{i-1} and, by induction hypothesis, the length of s_3 is 1.

Case 2: $k = 2$. Again, the proof is by induction on i and the induction base case $i = 0$ is trivially true, so we will suppose that the claim holds for $i-1$. By symmetry, we can further suppose without loss of generality that s_2 is a common subsequence of X_i and Y_i. If s_2 only uses labels from the image of one $r \in \{r_1, r_2, r_3\}$, then

$r^{-1}(s_2)$ is a common subsequence of X_{i-1} and Y_{i-1} so the claim holds by induction hypothesis. Otherwise, s_2 uses labels of at least two of r_1, r_2, and r_3. Since all labels of r_3 succeed all labels of r_1 and r_2 in X_i but preceed them in Y_i, we know that s_2 uses labels of r_1 and r_2 but not of r_3. Thus, s_2 admits two subsequences s' and s'' such that s' and s'' contain only labels of r_1 and r_2, respectively, and $|s'| + |s''| = |s_2|$. Then, however, s' is a subsequence of the result of removing all labels of r_2 and r_3 from X_i and Y_i, that is, of $r_1(X_{i-1})$ and $r_1(Y_{i-1})$ (see Construction 1). Thus, $r_1^{-1}(s')$ is a common subsequence of X_{i-1} and Y_{i-1} and, by induction hypothesis, $|s'| \leq 2^{i-1}$. An analogous argument shows that $|s''| \leq 2^{i-1}$ and, thus, $|s_2| = |s'| + |s''| \leq 2^i$. □

In the following, we prove lower bounds on the number of leaves in any MUL-tree displaying k of the caterpillars in $\mathcal{C}_i$ for $k \in \{1, 2, 3\}$ and $i \in \mathbb{N}$. We denote these bounds by $N_i^{(k)}$ and we note that, by the concatenation argument, $N_i^{(3)} \leq N_i^{(2)} + N_i^{(1)}$ and $N_i^{(2)} \leq 2N_i^{(1)}$ and $N_i^{(1)} = n(X_i) = n(Y_i) = n(Z_i) = 3^i$.

Lemma 1. *Let $i \in \mathbb{N}$ and let T be any MUL-tree displaying X_i and Y_i. Then, $n(T) \geq 2 \cdot 3^i - 2^i$.*

Proof. The proof is by induction on i. For the induction base, observe that all of X_0, Y_0 and T consist of a single leaf and $n(T) = 2 - 1 = 1$. For the induction step, suppose that the lemma holds for all $j < i$. Let T_A, T_B and T_C denote the subtrees of T induced by labels in A_i, B_i and C_i, respectively, and observe that their label multisets are a partition of the label multiset of T since A_i, B_i and C_i are disjoint. In the following, we show that at least one of T_A, T_B and T_C contains $2N_{i-1}^{(1)}$ leaves. Since, by definition, the other two contain at least $N_{i-1}^{(2)}$ leaves, we have

$$n(T) = n(T_A) + n(T_B) + n(T_C) \geq 2N_{i-1}^{(2)} + 2N_{i-1}^{(1)} \overset{\text{Ind. Hyp.}}{\geq} 2(2 \cdot 3^{i-1} - 2^{i-1}) + 2 \cdot 3^{i-1} = 2 \cdot 3^i - 2^i.$$

Towards a contradiction, assume that T_A, T_B and T_C contain less than $2N_{i-1}^{(1)}$ leaves. For all $F \in \{A_i, B_i, C_i\}$, let χ_F and ψ_F be respective embeddings of $X_i[F]$ and $Y_i[F]$ into T_F and note that, by assumption, T_F contains strictly less than $2N_{i-1}^{(1)} = n(X_i) + n(Y_i)$ leaves, implying that some leaf ℓ_F in T_F is mapped-to by both χ_F and ψ_F. In particular, the label of each ℓ_F occurs only once in T_F and, thus, in T, which allows us to use ℓ_F and its label interchangeably. Now, since $A_i <_{X_i} B_i <_{X_i} C_i$ we know that $\mathrm{LCA}(\ell_{A_i}, \ell_{C_i})$ is a strict ancestor of $\mathrm{LCA}(\ell_{A_i}, \ell_{B_i})$ in X_i and, since T displays X_i, this also holds in T. But since $C_i <_{Y_i} A_i <_{Y_i} B_i$ we also know that $\mathrm{LCA}(\ell_{A_i}, \ell_{B_i})$ is a strict ancestor of $\mathrm{LCA}(\ell_{A_i}, \ell_{C_i})$ in T, which is clearly a contradiction. □

Corollary 1. *Let $\epsilon > 0$. Then, there is some $n \in \mathbb{N}$ and two caterpillar trees of the same set of n labels, such that any MUL-tree displaying them has at least $(2 - \epsilon)n$ leaves.*

Proof. Let $i \in \mathbb{N}$ such that $(2/3)^i \leq \epsilon$ and, hence, $2^i \leq 3^i\epsilon = n\epsilon$. Let T be any MUL-tree displaying X_i and Y_i. Then, by Lemma 1, $n(T) \geq 2 \cdot 3^i - 2^i \geq 2n - n\epsilon = (2 - \epsilon)n$. □

Lemma 2. *Let $i \in \mathbb{N}$ and let T be any MUL-tree displaying X_i, Y_i and Z_i. Then, $n(T) \geq 3^{i+1} - 2^{i+1}$.*

Proof. The proof is by induction on i. For the induction base, observe that all of X_0, Y_0, Z_0, and T consist of a single leaf and $n(T) = 3 - 2 = 1$. For the induction step, suppose that the lemma holds for all $j < i$. Let T_A, T_B and T_C denote the subtrees of T induced by labels in A_i, B_i and C_i, respectively, and observe that their label multisets are a partition of the label multiset of T since A_i, B_i and C_i are disjoint. If any of T_A, T_B, and T_C contains $3N_{i-1}^{(1)} = 3^i$ leaves, then

$$n(T) = n(T_A) + n(T_B) + n(T_C) \geq 2N_{i-1}^{(3)} + 3N_{i-1}^{(1)} \overset{\text{Ind. Hyp.}}{\geq} 2(3^i - 2^i) + 3^i = 3^{i+1} - 2^{i+1}.$$

Further, if any two of T_A, T_B, and T_C contain $N_{i-1}^{(2)} + N_{i-1}^{(1)} \overset{\text{Lem 1}}{\geq} 2 \cdot 3^{i-1} - 2^{i-1} + 3^{i-1} = 3^i - 2^{i-1}$ leaves, then

$$n(T) = n(T_A) + n(T_B) + n(T_C) \geq N_{i-1}^{(3)} + 2(N_{i-1}^{(2)} + N_{i-1}^{(1)}) \overset{\text{Ind. Hyp.}}{\geq} 3^i - 2^i + 2(3^i - 2^{i-1}) = 3^{i+1} - 2^{i+1}$$

Thus, in the following, suppose that neither of the two cases holds. In particular, at least two trees among T_A, T_B, and T_C contain strictly less than $N_{i-1}^{(2)} + N_{i-1}^{(1)}$ leaves. By symmetry, suppose these are T_A and T_B.

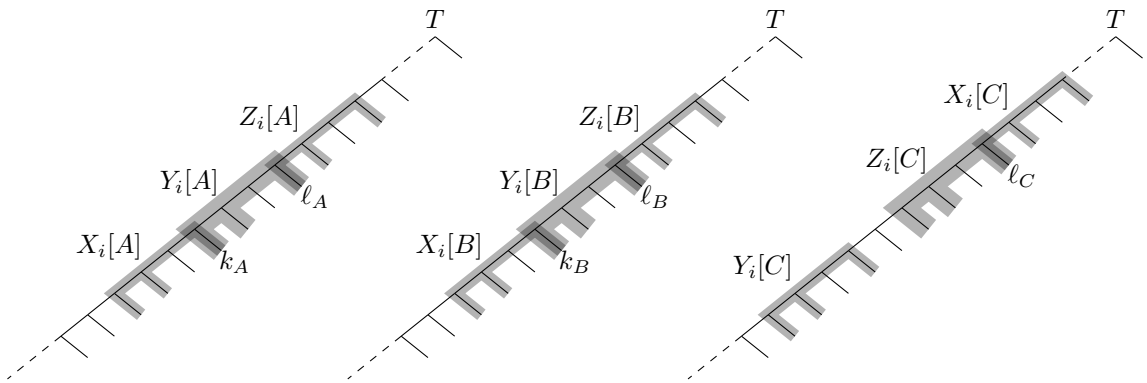

Figure 3. Illustration of the five leaves handled in the proof of Lemma 2. The three parts depict possible embeddings of $X_i[F]$, $Y_i[F]$, and $Z_i[F]$ for all $F \in \{A_i, B_i, C_i\}$. We assume that the embedding ψ_A of $Y_i[A_i]$ overlaps χ_A in k_A and ϕ_A in ℓ_A (left part) and, likewise for B. For C, only χ_C and ϕ_C overlap, in ℓ_C (right part). While T is not necessarily a caterpillar, drawing T linearly like that may help understand the situation.

For each $F \in \{A, B, C\}$, let χ_F, ψ_F, and ϕ_F denote the respective embeddings of $X_i[F_i]$, $Y_i[F_i]$ and $Z_i[F_i]$ into T_F and let us say that two among them *overlap* if T_F has a leaf that is assigned-to by both. Note that removing the $n(X_i[A_i]) = n(X_{i-1}) = N_{i-1}^{(1)}$ leaves of T_A that are mapped-to by χ results in a MUL-tree with strictly less than $N_{i-1}^{(2)}$ leaves and, by Lemma 1, this MUL-tree cannot display both $Y_i[A_i]$ and $Z_i[A_i]$). Thus, χ_A overlaps one of ψ_A and ϕ_A and the analog holds for ψ_A and ϕ_A. By pigeonhole principle, one among the three embeddings overlaps both others (while the other two may not necessarily overlap). Let ℓ_A and k_A denote the corresponding leaves (possibly $\ell_A = k_A$ if all three embeddings assign to ℓ_A). Since the same argument holds for B, we define ℓ_B and k_B analogously (see Fig. 3 for an illustration). Now, since T_C contains strictly less than $3N_{i-1}^{(1)}$ leaves by assumption, we know that two of χ_C, ψ_C, and ϕ_C overlap in a leaf ℓ_C of T_C. Next, we consider the relative positions of these five leaves in T.

In the following, a leaf a in T_A with parent t is said to be *above* a leaf b in T_B (written $a \rightsquigarrow b$) if $b <_T t$, and analogously for any pair chosen from the combined leaf-set of T_A, T_B and T_C. A third leaf ℓ is said to be *between* a and b if $a \rightsquigarrow \ell \rightsquigarrow b$. For all leaves a whose parent is an ancestor of b in T_A, we have that a is above both a and b and, likewise, for T_B and T_C. By symmetry, suppose that $\ell_F \rightsquigarrow k_F$ for all $F \in \{A, B\}$

Now, as $C <_{Y_i} A$ and $C <_{Z_i} A$, all leaves of ψ_C and ϕ_C are below all leaves of ψ_A and ϕ_A. Since ℓ_C is contained in the former and both ℓ_A and k_A are contained in the latter, we have that ℓ_A and k_A are both above ℓ_C. Further, as $B <_{X_i} C$ and $B <_{Z_i} C$, we have $\ell_C \rightsquigarrow \ell_B$. Thus, $\ell_A \rightsquigarrow k_A \rightsquigarrow \ell_C \rightsquigarrow \ell_B \rightsquigarrow k_B$, in particular both of ℓ_A, k_A are above both of ℓ_B, k_B. However, the caterpillar X_i contains at least one leaf mapped to ℓ_A or k_A, and at least one leaf mapped to ℓ_B or k_B. But since $A <_{X_i} B$, this implies that at least one of ℓ_B and k_B is above one of ℓ_A and k_A, contradicting $\ell_A \rightsquigarrow k_A \rightsquigarrow \ell_B \rightsquigarrow k_B$. □

Corollary 2. *Let $\epsilon > 0$. Then, there is some $n \in \mathbb{N}$ and three caterpillar trees of the same set of n labels, such that any MUL-tree displaying them has at least $(3 - \epsilon)n$ leaves.*

Proof. Let $i \in \mathbb{N}$ such that $(2/3)^i \leq \epsilon/2$ and, hence, $2^{i+1} \leq 3^i\epsilon = n\epsilon$. Let T be any MUL-tree displaying X_i, Y_i and Z_i. Then, by Lemma 2, $n(T) \geq 3^{i+1} - 2^{i+1} \geq 3n - n\epsilon = (3 - \epsilon)n$. □

4 Lower Bound on Networks

In this section, we build on the lower bound developed for MUL-trees in Section 3 to prove that, for any $\epsilon > 0$ and large enough n, any single-labeled phylogenetic network displaying the three n-leaf caterpillars constructed in Construction 1 has at least $(3/2 - \epsilon)n$ reticulations. To this end, we will give an algorithm that transforms any network displaying the caterpillars into a MUL-tree displaying the caterpillars by "unzipping" (or

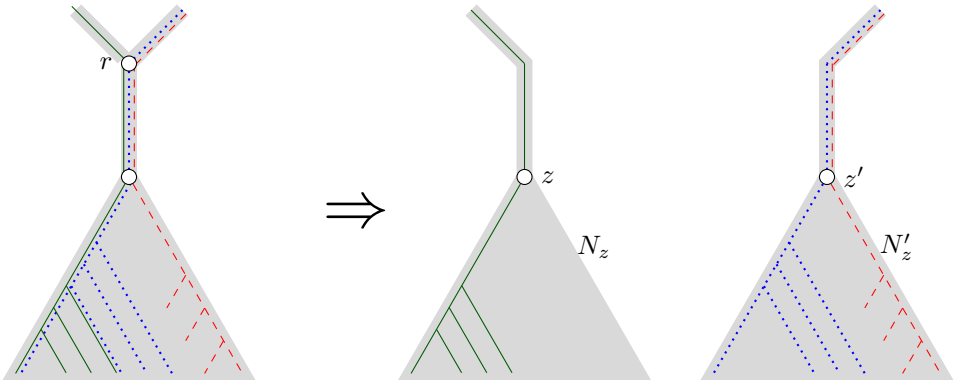

Figure 4. Illustration of the operation of "unzipping" (N, ϕ) at a lowest reticulation. The embedding of the three caterpillars X_i, Y_i and Z_i is depicted as green solid, red dashed and blue dotted lines, respectively, within the network outlined in gray. Note that all leaves of N_z into which ϕ embeds leaves of X_i as well as at least one of Y_i and Z_i, are duplicated in the process. Note also that not all three caterpillars are necessarily embedded in N_z, as previous unzip operations may have split a caterpillar off N_z.

"duplicating") subtrees. Then, we show that, if the network had fewer than $(3/2-\epsilon)n = (3/2-\epsilon)3^i$ reticulations, then the resulting MUL-tree has fewer than $3^{i+1} - 2^{i+1}$ leaves, contradicting Lemma 2.

4.1 Transforming the Network into a MUL-tree

In the following, we present a transformation acting on a given embedding of the three caterpillars into a multi-labeled network. Each application of our transformation rule will reduce the number of reticulations by one at the cost of creating new leaves. Hence, exhaustive application will result in an embedding of the three caterpillars into a MUL-tree. The rule acts on the subtree below a lowest reticulation and also manipulates a reservoir of virtual "tokens" which will help in the amortized analysis of how many new leaves are created in the process.

In the following, we work with pairs (N, ϕ), where N is a MUL-network with 3^i distinct labels and ϕ is an embedding of the caterpillars X_i, Y_i and Z_i (as constructed by Construction 1) into N such that all arcs of N are used by the embedding ϕ. We call such pairs *caterpillar embeddings*. Note that the assumptions that all arcs of N are used is satisfied by any embedding of the caterpillars into a network with smallest reticulation number. We make use of the fact that no embedding of any caterpillar can use both arcs incoming to any reticulation r of N, so the caterpillars with leaves embedded below r can be divided into two groups, depending on which incoming arc of r is used in their embedding. We call this the *parity* of a caterpillar with respect to r. We say that a caterpillar that does not have leaves below r has parity $\perp$ with respect to r. Note that, since all arcs of N are used by ϕ, there are two caterpillars with different non-$\perp$ parity with respect to r. Let N_r be the subnetwork of N rooted at r. We say that *the backbone of a caterpillar Q is embedded in N_r* (or *below r*) if ϕ maps a non-leaf of Q into N_r. Note that this is the case if and only if at least two leaves of N_r are used by the embedding of Q into N. The central operation in the transformations "unzips" (N, ϕ) at r (see Fig. 4).

Definition 1. *Let (N, ϕ) be a caterpillar-embedding, let r be a reticulation in N with child z such that the subnetwork N_z of N rooted at z does not contain reticulations. Let xr and yr denote the incoming arcs of r with $x \neq y$. The operation of* unzipping N at r *consists in the following steps:*

1. *Remove the node r from N.*
2. *Add a copy N'_z of N_z with root z' to N and add the arcs xz and yz'.*
3. *For each caterpillar Q such that ϕ embeds Q using the arc yr, replace all nodes u of N_z by their copy u' in N'_z in the embedding of Q.*
4. *repeatedly remove all leaves of the resulting MUL-network that are not used by the embedding, and suppress indegree-one outdegree-one nodes.*

Observation 1. *Let (N, ϕ) be a caterpillar embedding and let (N', ϕ') be the result of unzipping N at a reticulation r. Then, (N', ϕ') is a caterpillar embedding. In particular, all arcs of N' are used by ϕ'.*

The rest of this section depends on an arbitrary number $q \in \mathbb{N}$, which we will pick "sufficiently large" in the proof of the main theorem.[1] Further, we also suppose that our input caterpillars are "sufficiently large" with respect to q, that is, their length n satisfies

$$n > 12qn^{\log_3 2}. \tag{1}$$

Transformation Rule 1. *Let (N, ϕ) be a caterpillar-embedding, let r be a lowest reticulation in N, and let $\mathcal{Q}$ denote the set of caterpillars whose backbone is embedded in N_r. Then,*

1. *unzip N at r,*
2. *create three tokens in the token reservoir, and*
3. *for each leaf ℓ below r in N and each pair of different-parity caterpillars in $\mathcal{Q}$ whose embedding uses ℓ, remove $2q$ tokens from the token reservoir.*

As we will see, we never need to remove more tokens than are contained in the token reservoir.

Lemma 3. *Let N be a network with n leaves, let (N, ϕ) be a caterpillar embedding, let k be the number of reticulations of N, and let (T, ϕ') be the result of applying Transformation Rule 1 exhaustively to (N, ϕ). Then, T has at most $n + {}^{4(q+1)k}/_{3q}$ leaves.*

Proof. Intuitively, the proof is based on the observation that, whenever the transformation creates many new leaves for only a single reticulation it removes, then we can use half of these leaves to construct a common subsequence of two caterpillars. Then, Proposition 1 implies that this cannot happen too often.

Formally, we consider a series of "configurations" $C_0, C_1, \ldots, C_\Omega$, each consisting of a caterpillar embedding and a token reservoir where $C_0 := ((N, \phi), t_0 = 0)$ and $C_\Omega = ((T, \phi'), t_\Omega)$ for some t_Ω and each C_j results from an application of Transformation Rule 1 to the previous configuration C_{j-1}. To show Lemma 3, we assign a "weight" ω to each C_i. We prove that ω is monotonically non-increasing with respect to Transformation Rule 1. This will imply an upper bound on the number of leaves of the MUL-tree T displaying all three caterpillars. For a configuration $\mathcal{C} := ((\Gamma, \psi), t)$,

1. let $\#_{\mathrm{r}}^{\Gamma}$ denote the number of reticulations in Γ,
2. let $\#_i^{\Gamma}$ denote the number of leaves of Γ that are used by the embedding of exactly i caterpillars,

and define

$$\omega((\Gamma, \psi), t) := \sum_{i \in \{1,2,3,\mathrm{r}\}} c_i \cdot \#_i^{\Gamma} + c_t \cdot t \tag{2}$$

where $c_1 := c_3 := 3q$, $c_2 := 4q$, $c_r := 4(q+1)$, and $c_t := 1$. In the following, we omit the superscript Γ when it is clear from the context and we abbreviate the total number of leaves as $\#^{\Gamma} := \sum_{i \in \{1,2,3\}} \#_i^{\Gamma}$.

Claim 1. *ω is monotonically non-increasing with respect to Transformation Rule 1.*

Proof. We consider the following cases:

Case 1: No caterpillar has its backbone embedded in N_r. Then N_r has at most three leaves.

Case 1a: N_r contains two or three leaves, each with a single caterpillar embedded into it. Then, the numbers $\#_i$ do not change for any i, so ω increases by $\Delta\omega = 3c_t - c_r \leq 0$.

Case 1b: N_r contains two leaves, a leaf ℓ_1 with a single caterpillar embedded into it and a leaf ℓ_2 with two caterpillars embedded into it. If the two caterpillars whose embedding uses ℓ_2 have the same parity, then $\#_i$ does not change for any i, see Case 1a. Otherwise, $\#_1$ grows by two and $\#_2$ decreases by one, implying that ω grows by $\Delta\omega = (2c_1 + 3c_t) - (c_2 + c_r) = 6q + 3 - 8q - 4 \leq 0$.

Case 1c: N_r contains a single leaf ℓ with exactly two caterpillars embedded into it (their parity must differ in this case). Then, $\#_1$ grows by two and $\#_2$ decreases by one, implying that ω grows by $\Delta\omega = (2c_1 + 3c_t) - (c_2 + c_r) = 6q + 3 - 8q - 4 \leq 0$.

[1] We note that the proofs that follow also work for $q \in \mathbb{R}$, but for ease of presentation we assume $q \in \mathbb{N}$.

Case 1d: N_r contains a single leaf ℓ with three caterpillars embedded into it. Then, $\#_1$ and $\#_2$ each grow by one and $\#_3$ decreases by one, implying that ω grows by $\Delta\omega = (c_1+c_2+3c_t)-(c_3+c_r) = 7q+3-7q-4 \leq 0$.

Case 2: Exactly one caterpillar Q has its backbone embedded in N_r. Let L_Q denote the set of all (at least two) leaves below r that leaves of Q are embedded into.

Case 2a: All caterpillars with a leaf embedded into a leaf of L_Q have the same parity as Q (note that this case applies regardless of the parity of any caterpillar whose leaves are not embedded into any leaf of L_Q, even if they have a leaf embedded below r). Then, the numbers $\#_i$ do not change for any i, so ω increases by $\Delta\omega = 3c_t - c_r \leq 0$.

Case 2b: Exactly one leaf ℓ of L_Q is used to embed a leaf of a caterpillar with different parity than Q. If ℓ is used by exactly one caterpillar with different parity than Q, then $\#_1$ grows by two and $\#_2$ decreases by one, implying that ω grows by $\Delta\omega = (2c_1+3c_t)-(c_2+c_r) = 6q+3-8q-4 \leq 0$. If ℓ is used by all three caterpillars, at least one of which has different parity than Q, then $\#_1$ and $\#_2$ grow by one and $\#_3$ decreases by one, implying that ω grows by $\Delta\omega = (c_1+c_2+3c_t)-(c_3+c_r) = 7q+3-7q-4 \leq 0$.

Case 2c: Two leaves ℓ and ℓ' of L_Q are used to embed a leaf of a caterpillar with different parity than Q. Then, $\#_1$ grows by four and $\#_2$ decreases by two, implying that ω grows by $\Delta\omega = (4c_1+3c_t)-(2c_2+c_r) = 12q+3-12q-4 \leq 0$.

Case 3: Exactly two caterpillars Q and Q' have their backbone embedded in N_r and **their parity is the same**. Let L_Q and $L_{Q'}$ denote the sets of leaves in N_r that leaves in Q and Q', respectively, are embedded into.

Case 3a: No leaf of the third caterpillar is embedded in any leaf in $L_Q \cup L_{Q'}$. Then, the numbers $\#_i$ do not change for any i, so ω increases by $\Delta\omega \leq 3c_t - c_r \leq 0$

Case 3b: Exactly one leaf of the third caterpillar is embedded in a leaf ℓ in $L_Q \cup L_{Q'}$. If $\ell \notin L_Q \cap L_{Q'}$, then $\#_1$ grows by two and $\#_2$ decreases by one, implying that ω grows by $\Delta\omega = (2c_1+3c_t)-(c_2+c_r) = 6q+3-8q-4 \leq 0$. If $\ell \in L_Q \cap L_{Q'}$, then $\#_1$ and $\#_2$ grow by one and $\#_3$ decreases by one, implying that ω grows by $\Delta\omega = (c_1+c_2+3c_t)-(c_3+c_r) = 7q+3-7q-4 \leq 0$.

Case 4: Exactly two caterpillars Q and Q' have their backbone embedded in N_r and **their parity is different**. Let L_Q and $L_{Q'}$ denote the sets of leaves in N_r that leaves in Q and Q', respectively, are embedded into. Further, let $m := |L_Q \cap L_{Q'}|$.

Case 4a: The embedding of the third caterpillar uses no leaf in $L_Q \cup L_{Q'}$. Then $\#_1$ grows by $2m$, $\#_2$ decreases by m, and the token reservoir shrinks by $2qm-3$ tokens. Thus, ω grows by $\Delta\omega = (2mc_1+3c_t) - (mc_2+c_r+2mqc_t) = (6mq+3)-(4mq+4q+4+2mq) = -4q-1 \leq 0$.

Case 4b: The embedding of the third caterpillar uses a leaf of $L_Q \cap L_{Q'}$. Then, $\#_1$ grows by $2m-1$, $\#_2$ decreases by $m-2$, $\#_3$ decreases by one, and the token reservoir shrinks by $2qm-3$ tokens. Thus, ω grows by $\Delta\omega = ((2m-1)c_1+3c_t)-((m-2)c_2+c_3+c_r+2mqc_t) = (6mq-3q+3)-(4mq-8q+3q+4q+4+2mq) = -2q-1 \leq 0$.

Case 4c: The embedding of the third caterpillar uses a leaf of $L_Q \setminus L_{Q'}$. If the third caterpillar has the same parity as Q, then this is identical to Case 4a. Otherwise, $\#_1$ grows by $2(m+1)$, $\#_2$ decreases by $m+1$, and the token reservoir shrinks by $2qm-3$ tokens. Thus, ω grows by $\Delta\omega = (2(m+1)c_1+3c_t)-((m+1)c_2+c_r+2mqc_t) = (6mq+6q+3)-(4mq+4q+4q+4+2mq) = -2q-1 \leq 0$.

Case 4d: The embedding of the third caterpillar uses a leaf of $L_{Q'} \setminus L_Q$. This case is identical to Case 4c.

Case 5: All three caterpillars have their backbone embedded in N_r. Let L_2 be the set of leaves below r such that each leaf of L_2 is used by the embeddings of exactly two caterpillars and these caterpillars have different parity. Let L_3 be the set of leaves below r that are used in the embeddings of all three caterpillars, and observe that each such leaf causes us to remove $4q$ tokens from the reservoir. Further, abbreviate $m_2 := |L_2|$ and $m_3 := |L_3|$. Then, $\#_1$ grows by $2m_2+m_3$, $\#_2$ grows by m_3-m_2, $\#_3$ shrinks by m_3 and the token reservoir shrinks by $2qm_2+4qm_3-3$. Thus, ω grows by $\Delta\omega = ((2m_2+m_3)c_1+m_3c_2+3c_t)-(m_2c_2+m_3c_3+c_r+(2qm_2+4qm_3)c_t) = (6qm_2+7qm_3+3)-(6qm_2+7qm_3+4q+4) \leq 0$. ■

Claim 2. *For all configurations C_j, the token reservoir t_j is non-negative.*

Proof. We show that, when the first withdrawal from the token reservoir happens, the number of tokens accumulated in the reservoir exceeds the total number of tokens ever to be removed from it. To this end, consider what happens when Transformation Rule 1 is applied to a lowest reticulation r in N^j for some iteration $((N^j, \phi_j), t_j)$. Recall that tokens are removed only if, for some caterpillars P and Q whose backbones are embedded below r and that have different parity below r, P and Q share some leaf ℓ below r. In such a case, we call r *bad* with respect to (P, Q), and ℓ is called r*-bad* with respect to (P, Q) (we omit the prefix if r is unknown). Note that no leaf is r-bad with respect to (P, Q) for more than one r, since, after applying Transformation Rule 1 to r, all r-bad leaves are "unzipped" and no longer shared by P and Q.

Now, fix P and Q, and consider only those leaves and reticulations that are bad with respect to (P, Q). In the following, we simply refer to such leaves and reticulations as "bad". Note that, since the embeddings of P and Q are subgraphs of the network N and the backbones of P and Q are embedded below each bad reticulation, the ancestor relation between bad reticulations is the same in the embedding of P as in the embedding of Q (otherwise, N contains a cycle). Since P and Q are caterpillars, there is a unique linear ordering $r_0, r_1, \ldots, r_m$ of the bad reticulations such that r_{i+1} is an ancestor of r_i in both P and Q for all i.

In the following, we construct a common subsequence of P and Q containing at least half of all bad leaves, which will imply the claim through use of Proposition 1. To this end, for each i, let s_i denote the sequence of r_i-bad leaves in P. Recalling that the r_i occur on the backbone of Q in the same order as they do in P, it suffices to show that a subsequence of Q can be obtained from s_i by removing at most one leaf and retaining at least one leaf. To this end, we conduct a closer inspection of the configuration $C_j = ((N^j, \phi_j), t_j)$ in which Transformation Rule 1 is applied to r_i. Suppose that there are at least two r_i-bad leaves since, otherwise, s_i is already a subsequence of Q. Let u denote the lowest node of the tree $N^j_{r_i}$ that still has the backbones of both P and Q embedded in it. Clearly, all r_i-bad leaves that are not below u form a suffix of s_i that is also a subsequence of Q. By definition of u, it is not a leaf of N^j and u has children v_P and v_Q in N^j such that at most one leaf ℓ_Q below v_P is used in the embedding of Q and at most one leaf ℓ_P below v_Q is used in the embedding of P. But then, removing ℓ_P from s_i yields a subsequence of Q. Note that the removal of ℓ_P is necessary to form a subseqence of Q since the sequences corresponding to P and Q may disagree on the relative ordering of ℓ_P and ℓ_Q.

Now, the concatenation s of all subsequences of Q corresponding to the s_i is a common subsequence of P and Q and it contains at least half of all bad leaves (recall that we only remove a leaf from s_i if it contains at least two leaves). Then, by Proposition 1, the number of bad leaves is at most $2 \cdot 2^i = 2 \cdot 2^{\log_3 n} = 2 \cdot n^{\log_3 2}$, where $n = 3^i$ is the number of leaves in N. Summing over all three caterpillar pairs, we get an upper bound of $6n^{\log_3 2}$ bad leaves overall.

Next, we show that the first token retraction is preceeded by the creation of enough tokens to compensate for all retractions. To this end, consider the first configuration $C_j = ((N^j, \phi_j), t_j)$ such that applying Transformation Rule 1 to a reticulation r in N^j incurs a withdrawal from the token reservoir. In particular, this implies the existance of two different-parity caterpillars P and Q such that their backbone is embedded below r and both their embeddings use a common leaf ℓ below r. However, by construction of P and Q, none of the labels occuring in the lowest third of P occurs in the lowest third of Q and, thus, ℓ is preceeded by at least $n/3$ leaves in either P or Q; without loss of generality, suppose Q. Since the backbone of Q is embedded below r, there are at least $n/3$ leaves below r used by the embedding of Q, but not that of P. However, since all leaves in N are used by all three caterpillars, all these $n/3$ leaves were "unzipped" in previous operations.

Since C_j is the first configuration in which P and Q have different parity and "share" a leaf, we know that in all previous "unzip" operations, either (a) P and Q have the same parity or (b) the embeddings of P and Q share no leaves below the corresponding reticulation or (c) the embedding of at least one of P and Q uses only one leaf below the corresponding reticulation. Clearly, in cases (a) and (b), the unzip operation does not "separate P from Q" in any leaf, that is, the unzip operation does not reduce the number of leaves used by the embeddings of both P and Q. In case (c), each unzip operation can "separate P from Q" in at most one leaf, implying that C_j is preceeded by at least $n/3$ unzip operations, each creating 3 tokens in the reservoir. Thus,

by the time the first withdrawal is made from the reservoir, it contains at least n tokens which is sufficient to cover all withdrawals (recall that we withdraw $2q$ tokens for each of the at most $6n^{\log_3 2}$ bad leaves) since $n > 12qn^{\log_3 2}$ by (1). ■

With Claim 1 and Claim 2 we can now prove the bound on the number of leaves $\#^T$ in T in the number n of leaves in N and the number k of reticulations in N. To this end, note and recall that (a) $t_0 = 0$, (b) all leaves in N are used by the embeddings of all three caterpillars, and (c) T has no reticulations. Then,

$$\begin{aligned} c_1 \cdot \#^T &= c_1 \sum_{i\in\{1,2,3\}} \#_i^T \leq \sum_{i\in\{1,2,3,\mathrm{r}\}} c_i \cdot \#_i^T \overset{Claim\ 2}{\leq} \sum_{i\in\{1,2,3,\mathrm{r}\}} c_i \cdot \#_i^T + c_t \cdot t_\Omega \overset{(2)}{=} \omega(C_\Omega) \\ &\overset{Claim\ 1}{\leq} \omega(C_0) \overset{(2)}{=} \sum_{i\in\{1,2,3,\mathrm{r}\}} c_i \cdot \#_i^N + c_t \cdot t_0 = c_3 \cdot \#_3^N + c_r \cdot \#_\mathrm{r}^N = 3qn + 4(q+1)k \\ &= c_1(n + {}^{4(q+1)k}/_{3q}). \end{aligned}$$

Thus, T has at most $n + {}^{4(q+1)k}/_{3q}$ leaves. □

Lemma 3 tells us that if we can construct a network with few reticulations that displays our caterpillars, then we can also construct a MUL-tree with few leaves that displays our caterpillars. Since Corollary 1 says that such MUL-trees do not exist, we conclude that such networks do not exist.

Theorem 1. *Let $\epsilon > 0$. Then, there are three caterpillar trees, each with $n \in \mathbb{N}$ leaves, such that any network displaying all three caterpillars has at least $({}^3/_2 - \epsilon)n$ reticulations.*

Proof. Let $\delta := {}^{2\epsilon}/_3$ and choose q large enough so that $\beta := \delta - {}^1/_{q+1} > 0$. Finally, choose i such that $\beta \geq (1 - {}^1/_{q+1})({}^2/_3)^i$. Let N be a network displaying X_i, Y_i and Z_i, that is, there is a caterpillar embedding (N, ϕ). Let (T, ϕ) be the result of applying Transformation Rule 1 exhausively to (N, ϕ) and note that, by Observation 1, (T, ϕ) is a caterpillar embedding, that is, T displays X_i, Y_i, and Z_i. Further, by Lemma 2, T has at least $3^{i+1} - 2^{i+1}$ leaves. Let k denote the number of reticulations in N. Then,

$$3^{i+1} - 2^{i+1} \overset{Lemma\ 2}{\leq} \#_T \overset{Lemma\ 3}{\leq} n + {}^{4(q+1)k}/_{3q} = 3^i + {}^{4(q+1)}/_{3q} \cdot k$$

and, thus,

$$\begin{aligned} k \geq (2 \cdot 3^i - 2^{i+1}) \cdot {}^{3q}/_{4(q+1)} &= {}^3/_2 \cdot {}^q/_{q+1} \cdot (n - 2^i) \\ &= {}^3/_2(1 - {}^1/_{q+1})(n - 2^i) \\ &\geq {}^3/_2((1 - {}^1/_{q+1})n - \beta n) \\ &= {}^3/_2(1 - \delta)n = ({}^3/_2 - \epsilon)n. \end{aligned}$$ □

5 Discussion and Open Problems

We have shown that, for each $\epsilon > 0$ and $t \leq 3$, there is some $n \in \mathbb{N}$, and t caterpillars with n leaves, such that any MUL-tree displaying the caterpillars has at least $(t-\epsilon)n$ leaves. Whether this result can be generalized to $t \geq 4$ remains an interesting open question, even more so for networks, where the question would be to generalize our lower bound of $({}^3/_2-\epsilon)n$ reticulations required to display three caterpillars to $t \geq 4$ caterpillars.

Note that Theorem 1 can be stated more precisely as "there is some function $\alpha(n) \in o(1)$ such that, for each n, there are three caterpillars with n leaves that cannot be displayed by any network with fewer than $({}^3/_2 - \alpha(n))n$ reticulations". This raises the question how the "smallest" functions α for which this statement still holds, may look like (other than being in $o(1)$). In the MUL-tree case, a closer inspection of our proofs gives $\alpha(n) \leq 2 \cdot ({}^2/_3)^{\log_3 n}$. If we force the MUL-tree displaying the t trees to be a caterpillar, the question is equivalent to the question of the shortest supersequence that any collection of t permutations over $\{1, \ldots, n\}$ may have. This can be shown [Hun23] to be at least $(t - O(n^{-1/2}))n$, implying $\alpha(n) \in O(n^{-1/2}) \subset o(({}^2/_3)^{\log_3 n})$.

If there is always a MUL-caterpillar among the optimal MUL-trees displaying the t caterpillars, then this would imply a stricter bound for both MUL-trees and networks than what we showed here. Apart from improving the function $\alpha(n)$, it is interesting whether the factor of $3/2$ can be improved. In other words, is there a family of triples of phylogenetic trees for which more than $3n/2$ reticulations are needed? We do know that all triples of phylogenetic trees can have a network with $2(n-2)$ reticulations that display them: the network can be obtained from a tree on two leaves by inserting each remaining leaf using two reticulations. However, the same three trees might also be displayed by a network with strictly fewer reticulations. The best possible bound for three trees is therefore between $(3/2-\epsilon)n$ and $2(n-2)$. More generally, the best possible bound for $t \geq 3$ trees is between $(3/2-\epsilon)n$ and $(t-1)(n-2)$. Can this gap be closed or narrowed? As our results are for caterpillars, we expect that trees of varying topology can be used to prove such tighter bounds.

The last two questions beg the, somewhat philosophical question of whether reticulations are strictly more powerful than multiple leaves? For MUL-trees, we know that we cannot do much better than the trivial upper bound of $t \cdot n$. Is the same true for networks, can we not do much better than the trivial upper bound of $(t-1)(n-2)$, or are networks really more powerful than MUL-trees in this sense?

Finally, our original motivation for considering this problem came from the Hybridization Number problem. However, the bounds proven in this paper do not have direct formal consequences for that problem. Hence, another interesting direction for future research is to see if our bounds can be used to prove a negative result regarding exact algorithms for Hybridization Number, e.g. parameterized by treewidth (see [IJW22]).

Acknowledgements

We thank Simone Linz, Guillaume Scholz, and Stefan Grünewald for their thorough review work and helpful comments. We also thank Zach Hunter for explaining the bound on supersequences of permutations.

Fundings

Leo van Iersel is funded in part by Netherlands Organization for Scientific Research (NWO) grants OCENW.KLEIN.125 and OCENW.GROOT.2019.015. Mark Jones is funded by Netherlands Organization for Scientific Research (NWO) grant OCENW.KLEIN.125.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article. All authors are also recommenders of PCI Mathematical & Computational Biology.

References

[BS07] Bordewich M and Semple C. Computing the minimum number of hybridization events for a consistent evolutionary history. *Discrete Applied Mathematics* 155 (2007), 914–928.

[BS18] Bordewich M and Semple C. A universal tree-based network with the minimum number of reticulations. *Discrete Applied Mathematics* 250 (2018), 357–362. issn: 0166-218X. https://doi.org/10.1016/j.dam.2018.05.010.

[BSS05] Baroni M, Semple C, and Steel M. A framework for representing reticulate evolution. *Annals of Combinatorics* 8 (2005), 391–408. https://doi.org/10.1007/s00026-004-0228-0.

[Bar+05] Baroni M, Grünewald S, Moulton V, and Semple C. Bounding the number of hybridisation events for a consistent evolutionary history. *Journal of mathematical biology* 51 (2005), 171–182.

[DARM08] Dagan T, Artzy-Randrup Y, and Martin W. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences* 105 (2008), 10039–10044.

[FS15] Francis AR and Steel M. Which Phylogenetic Networks are Merely Trees with Additional Arcs? *Systematic Biology* 64 (June 2015), 768–777. issn: 1063-5157. https://doi.org/10.1093/sysbio/syv037.

[Hay16] Hayamizu M. On the existence of infinitely many universal tree-based networks. *Journal of Theoretical Biology* 396 (2016), 204–206. issn: 0022-5193. https://doi.org/10.1016/j.jtbi.2016.02.023.

[Hun23] Hunter Z. *Shortest Common Supersequence of Permutations – TCS StackExchange*. 2023. url: https://cstheory.stackexchange.com/questions/53153 (visited on 08/06/2023).

[IJW22] Iersel L van, Jones M, and Weller M. Embedding Phylogenetic Trees in Networks of Low Treewidth. In: *30th Annual European Symposium on Algorithms (ESA 2022)*. Vol. 244. Leibniz International Proceedings in Informatics (LIPIcs). 2022, 69:1–69:14. isbn: 978-3-95977-247-1. https://doi.org/10.4230/LIPIcs.ESA.2022.69.

[IK11] Iersel L van and Kelk S. When two trees go to war. *Journal of theoretical biology* 269 (2011), 245–255.

[Kel+12] Kelk S, Iersel L van, Lekic N, Linz S, Scornavacca C, and Stougie L. Cycle killer... qu'est-ce que c'est? On the comparative approximability of hybridization number and directed feedback vertex set. *SIAM journal on discrete mathematics* 26 (2012), 1635–1656.

[Mal05] Mallet J. Hybridization as an invasion of the genome. *Trends in ecology & evolution* 20 (2005), 229–237.

[Zha16] Zhang L. On tree-based phylogenetic networks. *Journal of Computational Biology* 23 (2016), 553–565.