ERNST MORITZ ARNDT
UNIVERSITÄT GREIFSWALD

Wissen
lockt.
Seit 1456

Extending the concept of phylogenetic diversity and its measures from trees to networks

MASTER'S THESIS

submitted in partial fulfillment of the requirements

for the degree of Master of Science (M.Sc.) in Biomathematics by

Kristina Wicke

Institute of Mathematics and Computer Science

Ernst-Moritz-Arndt-Universität Greifswald

First supervisor: Prof. Dr. Mareike Fischer

Second supervisor: Prof. Dr. Volkmar Liebscher

Greifswald, 8th November 2016

Abstract

In biodiversity conservation, it is often necessary to prioritize the species to conserve. Existing approaches to prioritization, e.g. the *Fair Proportion Index* and the *Shapley Value*, are based on phylogenetic trees and rank species according to their contribution to overall *phylogenetic diversity*. However, in many cases evolution is not treelike and thus, phylogenetic networks have come to the fore as a generalization of phylogenetic trees, allowing for the representation of non-treelike evolutionary events, such as horizontal gene transfer or hybridization.

In this thesis we extend the concept of *phylogenetic diversity* and its measures from phylogenetic trees to phylogenetic networks, in particular to hybridization networks. On the one hand, we consider the treelike content of a phylogenetic network, e.g. the (multi)set of phylogenetic trees displayed by a network and the *LSA tree* associated with it. On the other hand, we derive the *phylogenetic diversity* of subsets of taxa and biodiversity indices directly from the internal structure of a phylogenetic network. Furthermore, we introduce a small program that allows for the calculation of *phylogenetic diversity* and biodiversity indices based on phylogenetic networks. We illustrate some of the approaches using data for a group of marine mammals, the family Delphinidae, whose evolution is suspected to have included hybridization.

In summary, our approaches are an extension of existing prioritization tools in conservation biology and allow for the consideration of phylogenetic networks in prioritization decisions.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

DAG ..... Directed Acyclic Digraph
ED ..... Evolutionary Distinctiveness
EDGE ..... Evolutionary Distinct and Globally Endangered
FP ..... Fair Proportion Index
$\mathbf{FP}^{*}_{\mathsf{T}(\mathcal{N})}$ ..... embedded Fair Proportion Index
$\mathbf{FP}^{\mathbf{LSA}_*}$ ..... LSA associated Fair Proportion Index
HGT ..... Horizontal Gene Transfer
IPD ..... Inherited Phylogenetic Diversity
LCA ..... Lowest Common Ancestor
LGT ..... Lateral Gene Transfer
LSA ..... Lowest Stable Ancestor
NFP ..... Net Fair Proportion Index
PD ..... Phylogenetic Diversity
$\mathbf{PD}^{*}_{\mathsf{T}(\mathcal{N})}$ ..... embedded Phylogenetic Diversity
$\mathbf{PD}^{\mathbf{LSA}}$ ..... LSA associated Phylogenetic Diversity
PND ..... Phylogenetic Net Diversity
$\mathbf{PND}^{\mathbf{hyb}}$ ..... Hybrid Phylogenetic Net Diversity
$\mathbf{PND}^{\mathbf{ML}}$ ..... Maximum Likelihood Phylogenetic Net Diversity
SV ..... Shapley Value
$\widetilde{\mathbf{SV}}$ ..... modified Shapley Value
$\widehat{\mathbf{SV}}$ ..... unrooted rooted Shapley Value
$\mathbf{SV}^{*}_{\mathsf{T}(\mathcal{N})}$ ..... embedded Shapley Value
$\widetilde{\mathbf{SV}}^{*}_{\mathsf{T}(\mathcal{N})}$ ..... embedded modified Shapley Value
$\widehat{\mathbf{SV}}^{*}_{\mathsf{T}(\mathcal{N})}$ ..... embedded unrooted rooted Shapley Value
$\mathbf{SV}^{\mathbf{LSA}_*}$ ..... LSA associated Shapley Value
$\widetilde{\mathbf{SV}}^{\mathbf{LSA}_*}$ ..... LSA associated modified Shapley Value
$\widehat{\mathbf{SV}}^{\mathbf{LSA}_*}$ ..... LSA associated unrooted rooted Shapley Value
$\mathbf{SV}_{\mathcal{PD}}$ ..... generalized Shapley Value
$\widetilde{\mathbf{SV}}_{\mathcal{PD}}$ ..... generalized modified Shapley Value
TCTC ..... Tree-Child and Time-Consistent
$\mathcal{T}_{\mathbf{LSA}}$ ..... LSA Tree
$\mathcal{T}^{\mathbf{hyb}}_{\mathbf{LSA}}$ ..... Hybrid LSA Tree
$\mathcal{T}^{\mathbf{ML}}_{\mathbf{LSA}}$ ..... Maximum Likelihood LSA Tree

1. Introduction

Facing a major extinction crisis and the inevitable loss of biodiversity at the same time with limited financial means, biological conservation has to prioritize the species to conserve. One major objective of conservation biology is to consider overall biodiversity and minimize its future loss. In this matter, the so-called *phylogenetic diversity* (cf. Faith [11]) has come to the fore, measuring biodiversity based on the evolutionary history of species. Given a phylogenetic tree, *phylogenetic diversity* captures the diversity within a set of species and serves as a basis for biodiversity indices used in taxon prioritization. Two biodiversity indices frequently discussed are the *Fair Proportion Index* and the *Shapley Value*, which rank species according to their contribution to overall *phylogenetic diversity* (cf. Haake et al. [16], Hartmann [18], Fuchs and Jin [15], Wicke and Fischer [37]) .

Both *phylogenetic diversity*, as well as the *Fair Proportion Index* and the *Shapley Value* are based on phylogenetic trees and thus, assume the evolutionary history of species to be treelike. However, there are several forms of non-treelike evolution, such as *horizontal gene transfer* or *hybridization*, affecting a variety of species. Therefore phylogenetic reticulation networks have become an important concept in evolutionary biology, allowing for the representation of non-treelike evolution.

In this thesis we aim at combining both approaches, i.e. we aim at extending the concept of *phylogenetic diversity* and its measures from phylogenetic trees to phylogenetic networks, in particular to hybridization networks. So far, *phylogenetic diversity* and the *Shapley Value* have been considered for so-called *split networks*, which can be used to represent conflict in data (cf. Minh et al. [27], Volkmann et al. [33]), but no attempts have been made towards the generalization of *phylogenetic diversity* and its measures to reticulation networks.

After recapitulating the concepts of *phylogenetic diversity*, the *Fair Proportion Index* as well as the different versions of the *Shapley Value* and a short introduction to phylogenetic reticulation networks, we will therefore suggest several approaches towards the generalization of *phylogenetic diversity* and its measures from trees to networks, focusing on hybridization networks.

We will introduce a variety of definitions for generalized *phylogenetic diversity*, following three main principles: the calculation of spanning arborescences in a network, the consideration of the (multi)set of phylogenetic trees displayed by a network and the construction of the so-called *LSA tree* associated with a network.

We will then turn our attention to the *Fair Proportion Index* and the different versions of the *Shapley Value* and suggest different ways of using them as taxon prioritiza-

tion tools in the context of hybridization networks. On the one hand, we will again use the (multi)set of phylogenetic trees displayed by a network and the *LSA tree* associated with a network when calculating the *Fair Proportion Index* and the different versions of the *Shapley Value* for the taxa of a phylogenetic network. On the other hand, we will additionally derive the *Shapley Value* from any measure of generalized *phylogenetic diversity* and introduce a new index – the *Net Fair Proportion Index* – very similar to the *Fair Proportion Index*, but defined for rooted phylogenetic networks.

Both for the generalized measures of *phylogenetic diversity* and the generalized biodiversity indices, we will develop approaches that are independent of hybridization probabilities and approaches that explicitly incorporate these probabilities. In case of the former, we will additionally introduce a small program, `net_diversity.pl`, that allows for the computation of generalized *phylogenetic diversity* and generalized biodiversity indices independent of hybridization probabilities.

We will conclude this thesis with the application of some of the discussed concepts, in particular the *Shapley Value*, to a phylogenetic hybridization network for dolphins from the family Delphinidae.

2. Phylogenetic diversity and biodiversity indices on trees

Phylogenetic diversity plays an important role in conservation biology. It aims at quantifying the evolutionary distinctiveness of a single species and capturing the diversity within a set of species. Based on *phylogenetic diversity*, several indices have been developed in order to prioritize the species to conserve.

Before we go into more detail about *phylogenetic diversity* and diversity indices, we need to introduce some notations and definitions.

2.1. Preliminaries

Primarily, we define what we recognize as a *directed path* throughout this thesis, before introducing *phylogenetic trees* and *arborescences*.

Definition 1 (Directed path)**.** Let $G = (V, E)$ be a directed graph with node set V and edge set E. A *directed path* $P = (v_1, \ldots, v_{k+1})$ from v_1 to v_{k+1} is a sequence of nodes from V that are distinct (except possibly for the first and last node), such that there exists a directed edge $e_i = (v_i, v_{i+1}) \in E$ for all $i = 1, \ldots, k$.
If $v_1 = v_{k+1}$ we call P a *directed cycle*.
We sometimes call P a v_1–v_{k+1}–path to emphasize the start and end node of P.
If there is a function $c : E \to \mathbb{R}$ that assigns a length to each edge in E, we define the *length* of the path P as

$$length(P) = \sum_{e \in P} c(e),$$

where the sum runs over all edges of P.

Remark. For convenience we sometimes speak of *paths* instead of *directed paths*.

Definition 2 (Phylogenetic X-tree)**.** Let $T = (V(T), E(T))$ be a (graph-theoretical) tree with nodes $V(T)$, edges $E(T)$ on a leaf set $V_L \subseteq V(T)$ and no nodes of degree 2. Let X be a set of taxa and let $\phi : X \to V_L$ be a bijective mapping from the set of taxa into the set of leaves of T (X is therefore sometimes called *leaf set*). Then $\mathcal{T} := (T, \phi)$ is called a *phylogenetic X-tree* with *treeshape/topology* T.
If all internal nodes are of degree 3, we call $\mathcal{T}$ a *binary phylogenetic X-tree*. If there is a specified root node ρ, $\mathcal{T}$ is called a *rooted phylogenetic X-tree*. Last but not least, we call $\mathcal{T}$ a *rooted binary phylogenetic X-tree*, if $\mathcal{T}$ contains a specified root node ρ with $\deg(\rho) = 2$[1] and all other internal nodes have degree 3.

[1] For rooted binary phylogenetic X-trees, one node of degree 2 (the root) is allowed.

In case the edges of T have edge lengths assigned to them, we denote the length of an edge $e \in E(T)$ as λ_e.

Remarks.

- For convenience we speak of *phylogenetic trees* instead of *phylogenetic X-trees* when the set of taxa is clearly specified or can be assumed to be $X = \{1, 2, \ldots, n\}$. Moreover, when we refer to phylogenetic trees, we always mean binary phylogenetic trees, if not stated otherwise.

- In case of rooted phylogenetic trees we consider the edges to be directed away from the root. Thus, formally the treeshape T is a rooted directed acyclic graph or, to be more precise, an *arborescence* (cf. Definition 3) rather than a tree. However, we omit arrowheads when drawing rooted phylogenetic trees (cf. Figure 1).

Definition 3 (Arborescence)**.** Let $G = (V, E)$ be a directed graph and let $\rho \in V$ be a specified root node (of indegree 0). Then G is an *arborescence* (rooted at ρ) if there is exactly one directed path from ρ to u for all nodes $u \in V \setminus \rho$.

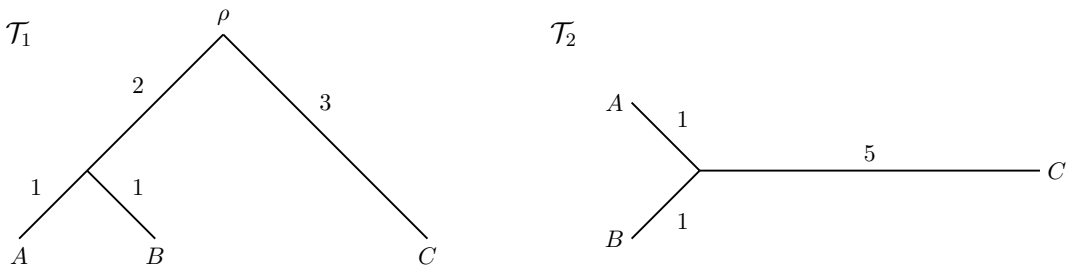


Fig. 1: Rooted binary phylogenetic X-tree $\mathcal{T}_1$ and unrooted binary phylogenetic X-tree $\mathcal{T}_2$ with leaf set $X = \{A, B, C\}$. Note that, formally, the edges in $\mathcal{T}_1$ are directed away from the root ρ, but for convenience arrowheads are omitted.

2.2. Phylogenetic diversity

Phylogenetic diversity, or *PD* for short, was first introduced by Faith [11] and has become an important measure of biodiversity. It captures the diversity within a set of species and serves as a basis for diversity indices used in taxon prioritization.

Mathematically, *phylogenetic diversity* is based on weighted phylogenetic trees, i.e. trees where the edges are assigned weights, representing for example time or substitution rates.

We are now in the position to formally define *phylogenetic diversity*, where we distinguish between rooted and unrooted phylogenetic trees.

Definition 4 (Phylogenetic diversity (PD))**.**

1. Let $\mathcal{T}_r$ be a rooted phylogenetic tree with leaf set X. For a subset $S \subseteq X$ of taxa the *phylogenetic diversity (PD)* of S is calculated by summing up the branch lengths of the phylogenetic subtree of $\mathcal{T}_r$ containing S and the root (i.e. we consider the sum of branch lengths in the smallest spanning tree containing S and the root).[2]

2. Now let $\mathcal{T}_u$ be an unrooted phylogenetic tree with leaf set X. Then the *phylogenetic diversity (PD)* of a subset $S \subseteq X$ of taxa is defined as the sum of branch lengths in the smallest spanning tree in $\mathcal{T}_u$ connecting those taxa. The PD of a single taxon is defined as 0.

Example 1. Consider the rooted phylogenetic tree $\mathcal{T}_1$ and the unrooted phylogenetic tree $\mathcal{T}_2$ depicted in Figure 1. We retrieve the following values for the *phylogenetic diversity* of $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively:

Rooted phylogenetic tree $\mathcal{T}_1$:

$$\begin{aligned}
PD_{\mathcal{T}_1}(\emptyset) &= 0,\\
PD_{\mathcal{T}_1}(\{A\}) &= 1+2=3,\\
PD_{\mathcal{T}_1}(\{B\}) &= 1+2=3,\\
PD_{\mathcal{T}_1}(\{C\}) &= 3,\\
PD_{\mathcal{T}_1}(\{A,B\}) &= 1+1+2=4,\\
PD_{\mathcal{T}_1}(\{A,C\}) &= 1+2+3=6,\\
PD_{\mathcal{T}_1}(\{B,C\}) &= 1+2+3=6,\\
PD_{\mathcal{T}_1}(\{A,B,C\}) &= 1+1+2+3=7.
\end{aligned}$$

Unrooted phylogenetic tree $\mathcal{T}_2$:

$$\begin{aligned}
PD_{\mathcal{T}_2}(\emptyset) = PD_{\mathcal{T}_2}(\{A\}) = PD_{\mathcal{T}_2}(\{B\}) = PD_{\mathcal{T}_2}(\{C\}) &= 0,\\
PD_{\mathcal{T}_2}(\{A,B\}) &= 1+1=2,\\
PD_{\mathcal{T}_2}(\{A,C\}) &= 1+5=6,\\
PD_{\mathcal{T}_2}(\{B,C\}) &= 1+5=6,\\
PD_{\mathcal{T}_2}(\{A,B,C\}) &= 1+2+5=7.
\end{aligned}$$

[2]Formally, we consider the smallest arborescence containing S and the root, because rooted phylogenetic trees are directed graphs.

Note that $\mathcal{T}_2$ can be obtained from $\mathcal{T}_1$ by suppressing the root node ρ, i.e. by deleting ρ and merging the two edges adjacent to ρ into one new branch, whereby adding the lengths of the former two branches to yield the length of the new branch.

However, unrooting a rooted phylogenetic tree causes a change in the definition of PD and thus, the values are in general not the same for the unrooted and rooted version of a tree. Consider for example $S = \{A, B\}$. For $\mathcal{T}_1$ we have $PD_{\mathcal{T}_1}(S) = 4$, while we have $PD_{\mathcal{T}_2}(S) = 2$ for $\mathcal{T}_2$. In this case, the difference between the two values can be explained by the edge of length 2 connecting S with the root in $\mathcal{T}_1$, which is disregarded in $\mathcal{T}_2$.

2.3. The Fair Proportion Index and the Shapley Value

The *Fair Proportion Index* and the *Shapley Value* have been frequently discussed as prioritization tools in biodiversity conservation (Haake et al. [16], Hartmann [18], Fuchs and Jin [15], Wicke and Fischer [37]). Both indices are based on phylogenetic trees and quantify the importance of a taxon to overall biodiversity. Thus, they provide a prioritization criterion to be used in conservation biology.

While the *Shapley Value* reflects the average biodiversity contribution of a species, the *Fair Proportion Index* lacks a biological link to conservation. It is, however, significantly easier to calculate and – under a different name (*ED* for *Evolutionary Distinctiveness*) – has been adopted to existing conservation schemes, such as the 'EDGE of Existence' project, established by the *Zoological Society of London* in 2007 (cf. Isaac et al. [22]). However, both indices have been shown to be highly correlated.

In the following, we will formally define the *Fair Proportion Index* and the *Shapley Value* and give an overview of the different definitions of the latter used in the literature.

2.3.1. The Fair Proportion Index

The *Fair Proportion Index*, or *FP* for short, is only defined for rooted phylogenetic trees. Its idea is to apportion the *phylogenetic diversity* of a tree among its leaves. This is achieved by distributing the length of each edge equally among the taxa descending from that edge.

Definition 5 (Fair Proportion Index (FP))**.** For a rooted phylogenetic tree $\mathcal{T}$ with leaf set X the *Fair Proportion Index* of a taxon a is defined as

$$FP(a) = \sum_{e} \frac{\lambda_e}{D_e}, \tag{2.1}$$

where the sum runs over all edges e on the path from the root to a and D_e denotes the number of leaves descendent from that edge.

Note that the sum of all *Fair Proportion Indices* for the taxa in X equals the total branch length of the given tree, since all edge weights are distributed equally among the descending taxa.

Example 2. For $\mathcal{T}_1$ in Figure 1 the *Fair Proportion Indices* are calculated as follows:

$$\begin{aligned} FP(A) &= \frac{2}{2} + 1 = 2, \\ FP(B) &= \frac{2}{2} + 1 = 2, \\ FP(C) &= 3. \end{aligned}$$

Summing up the *Fair Proportion Indices* we have $FP(A) + FP(B) + FP(C) = 7$, which equals the total sum of branch lengths in $\mathcal{T}_1$ or, in other words, the total PD (cf. Example 1).

2.3.2. The Shapley Value and its different versions

The original Shapley Value

The *Shapley Value* is used in different versions in the literature, namely the *original Shapley Value* SV, the *modified Shapley Value* $\widetilde{SV}$ and the *unrooted rooted Shapley Value* $\widehat{SV}$ (for an overview cf. Wicke and Fischer [37]).

The *original Shapley Value* SV was first introduced by Haake et al. [16] for unrooted phylogenetic trees, but can similarly be defined for rooted phylogenetic trees.

Definition 6 (Original Shapley Value (SV))**.** Let $\mathcal{T}$ be a phylogenetic tree with leaf set X and let $PD(S)$ denote the *phylogenetic diversity* of $S \subseteq X$. Then the *original Shapley Value* of a taxon a is defined as

$$SV(a) = \frac{1}{n!} \sum_{\substack{S \subseteq X \\ a \in S}} \Big((|S| - 1)!(n - |S|)!(PD(S) - PD(S \setminus \{a\})) \Big), \tag{2.2}$$

where $n = |X|$ and S denotes a subset of species containing taxon a (sometimes called a 'coalition' of taxa) and the sum runs over all such subsets possible.

Note that this general definition holds both for unrooted and rooted phylogenetic trees. The only difference is the way how *phylogenetic diversity* of subsets of taxa is defined (cf. Section 2.2).

For rooted trees, however, the *original Shapley Value* coincides with the *Fair Proportion Index*, which was recently shown by Fuchs and Jin [15].

In particular, the sum of the *original Shapley Values* for the taxa in X equals the total branch length of the given tree, just as it was the case for the *Fair Proportion Indices*. This property of the *original Shapley Value* is also referred to as the *Efficiency Axiom*. However, the *original Shapley Value* fulfills three other axioms, namely *Symmetry*, the *Dummy Axiom* and *Additivity* (cf. Haake et al. [16], Wicke [36] for details of the axioms and their meaning).

We now give an example for the calculation of the *original Shapley Value*.

Example 3. Consider the phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ in Figure 1.

We start with $\mathcal{T}_1$ and calculate the *original Shapley Value* for taxon A. Note that we have to consider 4 summands, as there are 4 subsets $S \subseteq X$ containing taxon A, namely $\{A\}$, $\{A, B\}$, $\{A, C\}$ and $\{A, B, C\}$.

$$\begin{aligned}
SV_{\mathcal{T}_1}(A) &= \frac{1}{3!} \sum_{S:A\in S} \Big((|S|-1)!(|X|-|S|)!(PD(S) - PD(S \setminus \{A\})) \Big) \\
&= \frac{1}{3!} \Big[(1-1)!(3-1)!(3-0) \\
&\quad + (2-1)!(3-2)!\big((4-3)+(6-3)\big) \\
&\quad + (3-1)!(3-3)!(7-6) \Big] \\
&= \frac{1}{6} \Big[1 \cdot 2 \cdot 3 + 1 \cdot 1 \cdot (1+3) + 2 \cdot 1 \cdot 1 \Big] \\
&= \frac{1}{6} \cdot 12 \\
&= 2.
\end{aligned}$$

Analogously, we get $SV_{\mathcal{T}_1}(B) = 2$ and $SV_{\mathcal{T}_1}(C) = 3$.

Note that the calculation of the *original Shapley Values* is much more involved than the calculation of the *Fair Proportion Indices* in Example 2, but as shown by Fuchs and Jin [15] the two values coincide for rooted trees.

We now turn our attention to the unrooted tree $\mathcal{T}_2$ and again, calculate the *original Shapley Value* for taxon A. The calculation is identical to the one for $\mathcal{T}_1$, the only difference being that we have to use other values for the *phylogenetic diversity* of

subsets of taxa (cf. Example 1). Thus, we get

$$\begin{aligned}
SV_{\mathcal{T}_2}(A) &= \frac{1}{3!}\sum_{S:A\in S}\Big((|S|-1)!(|X|-|S|)!(PD(S)-PD(S\setminus\{A\}))\Big)\\
&= \frac{1}{3!}\Big[(1-1)!(3-1)!(0-0)\\
&\quad + (2-1)!(3-2)!\big((2-0)+(6-0)\big)\\
&\quad + (3-1)!(3-3)!(7-6)\Big]\\
&= \frac{1}{6}\Big[1\cdot 2\cdot 0 + 1\cdot 1\cdot (2+6) + 2\cdot 1\cdot 1\Big]\\
&= \frac{1}{6}\cdot 10\\
&= \frac{5}{3}.
\end{aligned}$$

Analogously, $SV_{\mathcal{T}_2}(B) = \frac{5}{3}$ and $SV_{\mathcal{T}_2}(C) = \frac{11}{3}$.

Note that both for $\mathcal{T}_1$ and $\mathcal{T}_2$ the sum of the *original Shapley Values* equals the sum of the branch lengths of $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively. As mentioned in Example 1, $\mathcal{T}_2$ represents the unrooted version of $\mathcal{T}_1$, but due to the fact that unrooting a rooted tree causes a change in the definition of *phylogenetic diversity*, the *original Shapley Values* differ between $\mathcal{T}_1$ and $\mathcal{T}_2$.

The modified Shapley Value

We now turn our attention to a slightly modified version of the *Shapley Value* introduced by Fuchs and Jin [15], namely the *modified Shapley Value*, which we will denote as $\widetilde{SV}$. While the *original Shapley Value* considers all subsets of taxa containing a certain taxon, the *modified Shapley Value* only takes into account subsets of size at least 2. Again, it can be generally defined for both rooted and unrooted phylogenetic trees.

Definition 7 (Modified Shapley Value ($\widetilde{SV}$))**.** Let $\mathcal{T}$ be a phylogenetic tree with leaf set X and let $PD(S)$ denote the *phylogenetic diversity* of a subset $S \subseteq X$. Then the *modified Shapley Value* of a taxon a is defined as

$$\widetilde{SV}(a) = \frac{1}{n!}\sum_{\substack{S:a\in S\\ |S|\geq 2}}\Big((|S|-1)!(n-|S|)!(PD(S)-PD(S\setminus\{a\}))\Big), \tag{2.3}$$

where $n = |X|$ and the sum runs over all coalitions S containing taxon a and at least one other taxon.

Comparing the *original* and the *modified Shapley Value*, we get the following relation:

Proposition 1. *Let $\mathcal{T}$ be a phylogenetic tree with leaf set X and let $a \in X$ be a taxon of $\mathcal{T}$.*

1. *If $\mathcal{T}$ is rooted, we have*

$$SV(a) = \widetilde{SV}(a) + \frac{PD(\{a\})}{n}. \tag{2.4}$$

2. *If $\mathcal{T}$ is unrooted, the* original *and the* modified Shapley Value *coincide, thus, $SV(a) = \widetilde{SV}(a)$.*

Proof.

1. While the *original Shapley Value* considers all subsets of taxa containing taxon a, the *modified Shapley Value* only takes into account subsets of size at least two. Thus, we have to consider the contribution of the singleton set $\{a\}$ to $SV(a)$. By definition this is

$$\begin{aligned}\frac{1}{n!}\Big((1-1)!(n-1)!(PD(\{a\}) - PD(\emptyset))\Big) &= \frac{(n-1)!}{n!}PD(\{a\}) \\ &= \frac{PD(\{a\})}{n}.\end{aligned}$$

As $\widetilde{SV}(a)$ lacks this contribution, it is

$$\widetilde{SV}(a) = SV(a) - \frac{PD(\{a\})}{n},$$

or, in other words,

$$SV(a) = \widetilde{SV}(a) + \frac{PD(\{a\})}{n}$$

(taken from Wicke [36]).

2. The equality for the two values in the unrooted case follows from the equation above and the fact, that the *phylogenetic diversity* of a set $S \subseteq X$ is defined as 0, whenever S contains only one element (cf. Definition 4, part two), thus, in our case $PD(\{a\}) = 0$. In particular,

$$SV(a) = \widetilde{SV}(a) + \frac{0}{n} = \widetilde{SV}(a).$$

□

Remark. Note that the *modified Shapley Value* does not fulfill the *Efficiency Axiom* in the rooted case, i.e. the sum of the *modified Shapley Values* for the taxa in X does not equal the total branch length of the given rooted phylogenetic tree. To see this let $\mathcal{T}$ be a phylogenetic tree with leaf set X, where $|X| = n$. Then

$$\begin{aligned}
\sum_{i=1}^{n} \widetilde{SV}(i) &= \sum_{i=1}^{n} \Big(SV(i) - \frac{PD(\{i\})}{n} \Big) \\
&= \sum_{i=1}^{n} SV(i) - \sum_{i=1}^{n} \frac{PD(\{i\})}{n} \\
&= PD(X) - \sum_{i=1}^{n} \underbrace{\frac{PD(\{i\})}{n}}_{\substack{PD(\{i\})\neq 0, \\ \text{because } \mathcal{T} \text{ is rooted.}}} \\
&\neq PD(X).
\end{aligned}$$

Thereby

$$\sum_{i=1}^{n} SV(i) = PD(X)$$

holds because of the *efficiency* of the *original Shapley Value* (taken from Wicke [36]).

Example 4. We now calculate the *modified Shapley Values* for the phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ depicted in Figure 1. In contrast to the calculations in Example 3, however, we only have to consider subsets $S \subseteq X$ of size at least 2 now. Thus, we consider the sets $\{A, B\}$, $\{A, C\}$ and $\{A, B, C\}$.

For the rooted phylogenetic tree $\mathcal{T}_1$ this leads to

$$\begin{aligned}
\widetilde{SV}_{\mathcal{T}_1}(A) &= \frac{1}{3!} \sum_{\substack{S: A \in S \\ |S| \geq 2}} \Big((|S| - 1)!(|X| - |S|)!(PD(S) - PD(S \setminus \{A\})) \Big) \\
&= \frac{1}{3!} \Big[(2-1)!(3-2)!\big((4-3) + (6-3)\big) \\
&\qquad + (3-1)!(3-3)!(7-6) \Big] \\
&= \frac{1}{6} \Big[1 \cdot 1 \cdot (1+3) + 2 \cdot 1 \cdot 1 \Big] \\
&= \frac{1}{6} \cdot 6 \\
&= 1,
\end{aligned}$$

$\widetilde{SV}_{\mathcal{T}_1}(B) = 1$ and $\widetilde{SV}_{\mathcal{T}_1}(C) = 2$.

Recall that $PD(\{A\}) = 3, PD(\{B\}) = 3$ and $PD(\{C\}) = 3$. Comparing the

modified Shapley Values with the *original Shapley Values*, calculated in Example 3, we have

$$\begin{aligned}
\widetilde{SV}_{\mathcal{T}_1}(A) &= 1 = 2 - \frac{3}{3} = SV_{\mathcal{T}_1}(A) - \frac{PD(\{A\})}{n}, \\
\widetilde{SV}_{\mathcal{T}_1}(B) &= 1 = 2 - \frac{3}{3} = SV_{\mathcal{T}_1}(B) - \frac{PD(\{B\})}{n}, \\
\widetilde{SV}_{\mathcal{T}_1}(C) &= 2 = 3 - \frac{3}{3} = SV_{\mathcal{T}_1}(C) - \frac{PD(\{C\})}{n}
\end{aligned}$$

as implied by Proposition 1.
Also note that $\widetilde{SV}_{\mathcal{T}_1}(A) + \widetilde{SV}_{\mathcal{T}_1}(B) + \widetilde{SV}_{\mathcal{T}_1}(C) = 4 \neq 7 = PD(\{ABC\}) = PD(X)$.

Analogously, we have for the unrooted phylogenetic tree $\mathcal{T}_2$

$$\begin{aligned}
\widetilde{SV}_{\mathcal{T}_2}(A) &= \frac{1}{3!} \sum_{\substack{S: A \in S \\ |S| \geq 2}} \Big((|S|-1)!(|X|-|S|)!(PD(S) - PD(S \setminus \{A\})) \Big) \\
&= \frac{1}{3!} \Big[(2-1)!(3-2)! \big((2-0) + (6-0) \big) \\
&\qquad + (3-1)!(3-3)!(7-6) \Big] \\
&= \frac{1}{6} \Big[1 \cdot 1 \cdot (2+6) + 2 \cdot 1 \cdot 1 \Big] \\
&= \frac{1}{6} \cdot 10 \\
&= \frac{5}{3},
\end{aligned}$$

$\widetilde{SV}_{\mathcal{T}_2}(B) = \frac{5}{3}$ and $\widetilde{SV}_{\mathcal{T}_2}(C) = \frac{11}{3}$.
Comparing the results with those in Example 3, we see that the *original* and the *modified Shapley Value*, in fact, coincide for unrooted phylogenetic trees.

The unrooted rooted Shapley Value

Hartmann [18] observes a strong correlation between the *Fair Proportion Index* and the *Shapley Value* for rooted phylogenetic trees, but he does not come to the conclusion that they are equal. Thus, Hartmann [18] cannot be using the *original Shapley Value*, since this value coincides with the *Fair Proportion Index* as shown by Fuchs and Jin [15].

Fuchs and Jin [15] therefore suggest that the *modified Shapley Value* was used in Hartmann [18]. However, as the definition of the *Shapley Value* in Hartmann [18] ranges over all subsets $S \subseteq X$ containing a certain taxon, and not only over those of

size at least 2, Wicke and Fischer [37] think that yet another version of the *Shapley Value*, which they call the *unrooted rooted Shapley Value* $\widehat{SV}$, is used in Hartmann [18].

For a rooted phylogenetic tree $\mathcal{T}_r$ the *unrooted rooted Shapley Value* of a taxon is defined as the *original Shapley Value* of this taxon on the corresponding unrooted tree, i.e. on the tree $\mathcal{T}_u$ that is derived from the original tree $\mathcal{T}_r$ by suppressing the root node (cf. Wicke and Fischer [37]).

Definition 8 (Unrooted rooted Shapley Value ($\widehat{SV}$))**.** Let $\mathcal{T}_r$ be a rooted phylogenetic tree with leaf set X. Then we retrieve the *unrooted rooted Shapley Value* of a taxon $a \in X$ as

$$\widehat{SV}_{\mathcal{T}_r}(a) = SV_{\mathcal{T}_u}(a), \tag{2.5}$$

where $SV_{\mathcal{T}_u}(a)$ is the *original Shapley Value* of a in the corresponding unrooted phylogenetic tree $\mathcal{T}_u$.

Example 5. Consider the phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$ depicted in Figure 1. As explained in Example 1, the tree $\mathcal{T}_2$ is the unrooted version of the tree $\mathcal{T}_1$. In Example 3 we have already calculated the *original Shapley Values* for all leaves of $\mathcal{T}_2$. Thus, we already know the *unrooted rooted Shapley Values* for the taxa in $\mathcal{T}_1$. To be precise, we have $\widehat{SV}_{\mathcal{T}_1}(A) = \frac{5}{3}$, $\widehat{SV}_{\mathcal{T}_1}(B) = \frac{5}{3}$ and $\widehat{SV}_{\mathcal{T}_1}(C) = \frac{11}{3}$.

Summary

We finish this section with a short summary (taken from Wicke and Fischer [37]):
Let $\mathcal{T}_r$ be a rooted phylogenetic tree with leaf set X ($|X| = n$) and let $\mathcal{T}_u$ be its unrooted version, i.e. the unrooted phylogenetic tree that is derived from $\mathcal{T}_r$ by suppressing the root node. Then, we have:

- $SV_{\mathcal{T}_r} = FP_{\mathcal{T}_r}$ (proven by Fuchs and Jin [15]),
- $SV_{\mathcal{T}_r} \neq \widetilde{SV}_{\mathcal{T}_r}$, but $\widetilde{SV}_{\mathcal{T}_r}(a) = SV_{\mathcal{T}_r}(a) + \frac{PD_r(\{a\})}{n}$ for all $a \in X$,
- $SV_{\mathcal{T}_u} = \widetilde{SV}_{\mathcal{T}_u}$,
- $SV_{\mathcal{T}_r} \neq \widehat{SV}_{\mathcal{T}_r}$, but the two values are strongly correlated (cf. Hartmann [18]),
- $SV_{\mathcal{T}_u} = \widehat{SV}_{\mathcal{T}_r}$ and
- $\widehat{SV}_{\mathcal{T}_r} \neq \widetilde{SV}_{\mathcal{T}_r}$.

Phylogenetic diversity and biodiversity indices for phylogenetic trees have been of considerable interest in recent times. At the same time, the use of phylogenetic networks

has gained popularity among evolutionary biologists, as these allow for the modeling not only of speciation events, but also of reticulate evolution, e.g. of *hybridization* or *horizontal gene transfer*. However, there has been no attempt to extend *phylogenetic diversity* and its measures to reticulate networks, so far.

In the following we will first introduce some basic definitions and notations concerning phylogenetic networks. We will then suggest some ways of how the concept of *phylogenetic diversity* and diversity indices could be applied to phylogenetic networks.

3. Phylogenetic networks

Phylogenetic networks have become an important concept in evolutionary biology. In contrast to phylogenetic trees, they can be used to model reticulate (non-treelike) evolutionary events, such as *hybridization* or *horizontal gene transfer*.

Hybridization plays an important role in plant biology and 'is the process of interbreeding between individuals of different species (interspecific hybridization) or genetically divergent individuals from the same species (intraspecific hybridization)' (Hyb [1]).

Horizontal gene transfer (HGT), sometimes also referred to as *lateral gene transfer (LGT)*, on the other hand, describes the transmission of genetic material from one species to another. For *HGT* to take place, the two species have to coexist in time and thus, *HGT* differs from *vertical gene transfer*, where genetic material is passed from a parental organism to a child organism during reproduction. Horizontal gene transfer plays an important role in the evolution of prokaryotes, but can also be found in eukaryotes and is made possible by mobile DNA elements, such as plasmids, transposons or bacteriophages.

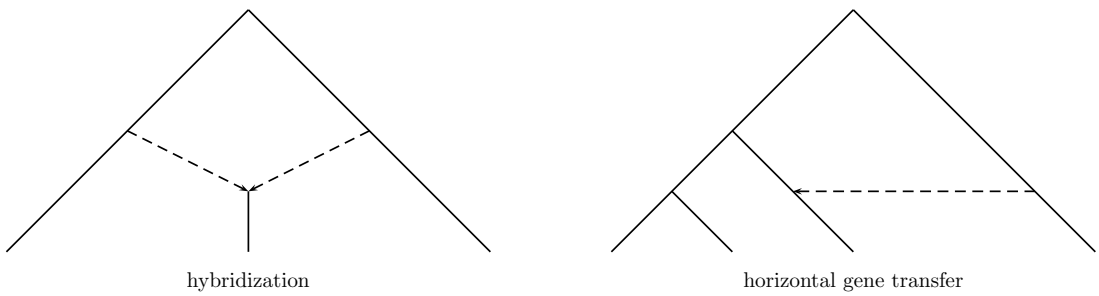


Fig. 2: Hybridization and horizontal gene transfer

Both processes, however, cannot be represented by phylogenetic trees. Therefore, phylogenetic networks were introduced as a mathematical generalization of phylogenetic trees. In the following, we will introduce some basic definitions and concepts concerning phylogenetic networks.

3.1. Definitions and notations

Definition 9 (Rooted binary phylogenetic network (cf. Cordue et al. [9]))**.** Let X be a set of taxa with $|X| = n$. A *rooted binary phylogenetic network* $\mathcal{N}$ on X is a connected, rooted acyclic digraph (rooted DAG) such that:

- the root has outdegree two (and indegree 0),
- each node with outdegree 0 has indegree 1, and the set of nodes with outdegree 0 is bijectively labeled by X,
- all other vertices either have indegree 1 and outdegree 2, or indegree 2 and outdegree 1.

Definition 10 (Types of nodes and edges in binary phylogenetic networks)**.** Let $\mathcal{N}$ be a phylogenetic network on some taxon set X (cf. Figure 3).

- The unique node with indegree 0 is called the *root* ρ.
- Nodes with indegree 2 and outdegree 1 are called *reticulation nodes* or *reticulations*, all other nodes are called *tree nodes*.
- A node with outdegree 0 is called a *leaf* node.
- Edges leading to a reticulation node are called *reticulation edges*, edges directed into a tree node are called *tree edges*.

Remarks.

- Note that a rooted binary phylogenetic X-tree $\mathcal{T}$ is a rooted binary phylogenetic network with no reticulation node.
- If $\mathcal{N}$ is a weighted DAG, we denote the length of an edge e as λ_e. Note, however, that in this thesis we only allow the tree edges of $\mathcal{N}$ to come with edge lengths, representing time or evolutionary change, while we consider the reticulation edges to be unweighted. W.l.o.g. we define the edge lengths of reticulation edges to be 0, if not stated otherwise. By doing so, functions on the edge lengths of a network, e.g. the sum of all edge lengths, are well-defined and we do not have to restrict them to tree edges in the following.
- When we refer to phylogenetic networks on X, we always mean rooted binary phylogenetic networks on X, if not stated otherwise.

In the following we will introduce three special cases of rooted phylogenetic networks, namely *tree-child* networks, *tree-sibling* networks and *time-consistent* networks, but to do so we need some more notations.

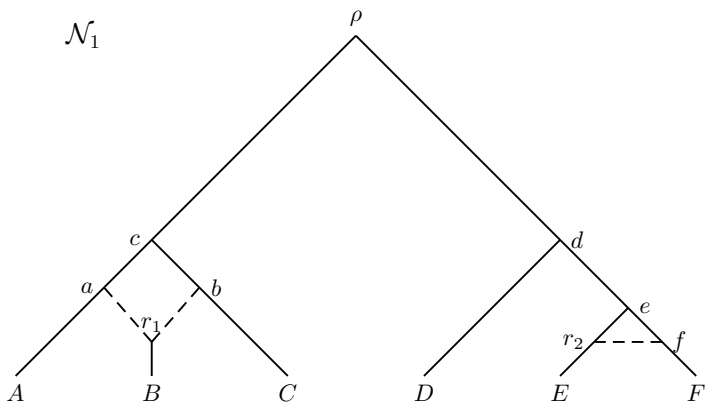


Fig. 3: Rooted binary phylogenetic network $\mathcal{N}_1$ on $X = \{A, B, C, D, E, F\}$ with *root node* ρ. There are two *reticulation nodes* r_1 and r_2, while all other nodes are *tree nodes*. Thus, the dashed edges directed into r_1 and r_2 are *reticulation edges* and all other edges are *tree edges*. The reticulation node r_1 can be interpreted as a hybridization event, the reticulation node r_2 as a horizontal gene transfer event. Note that, formally, all tree edges in $\mathcal{N}_1$ are directed away from the root ρ, but for convenience arrowheads are omitted. Similarily, all reticulation edges are directed into reticulation nodes, e.g. the edge (f, r_2) is directed into r_2.

Definition 11 (Ancestor, Descendant, Siblings (cf. Cordue et al. [9])). Let $\mathcal{N}$ be a phylogenetic network on some taxon set X and let u and v be two distinct nodes of $\mathcal{N}$. If there exists a directed path from u to v, we call u an *ancestor* of v and v is a *descendant* of u. If u and v are connected by an edge (i.e. the edge (u, v)), we say that u is a *parent* of v and v is a *child* of u. Two nodes that have a common parent are called *siblings*.

Example 6. Consider the phylogenetic network $\mathcal{N}_1$ with leaf set $X = \{A, B, C, D, E, F\}$ depicted in Figure 3. The root ρ is an *ancestor* of all other nodes and a *parent* of c and d. In other words, all nodes in $V(\mathcal{N}_1) \setminus \rho$ are *descendants* of ρ. In particular, c and d are *siblings* and *children* of ρ. Analogously, c is an *ancestor* of a, b, r_1, A, B and C, but only a *parent* of a and b. As a and b have a common parent, they are *siblings*. Easily, more examples of nodes that are *ancestors* or *descendants* of each other or that are *siblings* can be found.

We now introduce the so-called *tree-child property* of phylogenetic networks.

Definition 12 (Tree-child property, tree-child network (cf. Huson et al. [20], p. 164)). Let $\mathcal{N}$ be a rooted phylogenetic network on some taxon set X. A node v of $\mathcal{N}$ has the *tree-child property*, if it has at least one child u that is a tree node. We call $\mathcal{N}$ a *tree-child* network, if all internal nodes of $\mathcal{N}$ have the tree-child property (cf. Figure 4).

Less restrictive than the tree-child property is the so-called *tree-sibling property*.

Definition 13 (Tree-sibling property, tree-sibling network (cf. Huson et al. [20], p. 166))**.** Let $\mathcal{N}$ be a phylogenetic network on X. A node v of $\mathcal{N}$ has the *tree-sibling* property, if it has at least one sibling u that is a tree node. We call $\mathcal{N}$ a *tree-sibling network*, if all reticulation nodes have the tree-sibling property (cf. Figure 4).

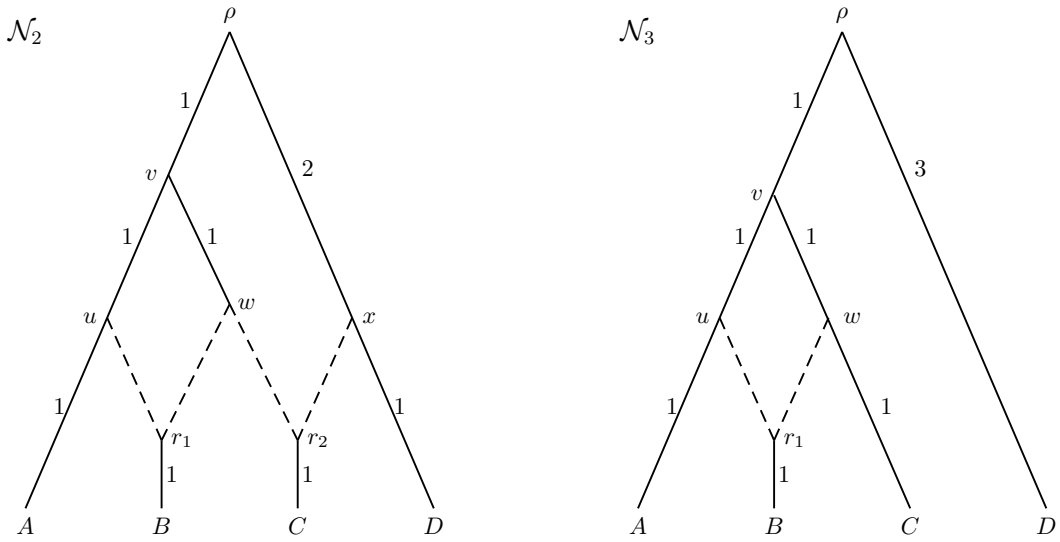


Fig. 4: Rooted binary phylogenetic networks $\mathcal{N}_2$ and $\mathcal{N}_3$ on $X = \{A, B, C, D\}$. $\mathcal{N}_2$ is a *tree-sibling* network, but not a *tree-child* network, because all children of the internal node w are reticulation nodes and thus, w has no children that are tree nodes. Still, all reticulation nodes in $\mathcal{N}_3$ have siblings that are tree nodes, and thus $\mathcal{N}_2$ is *tree-sibling*. The network $\mathcal{N}_3$ is both a *tree-sibling* and a *tree-child* network, because all internal nodes have at least one child that is a tree node.

When phylogenetic networks are used to model the evolution of some taxon set X, including hybridization or horizontal gene transfer events, we have to consider another topological constraint on the network: For a reticulation event to take place, and thus, for a hybrid species to arise, the two parent species have to coexist in time. This leads to the concept of *time-consistent* networks.

Definition 14 (Time-consistent network)**.** A phylogenetic network $\mathcal{N} = (V, E)$ with node set V and edge set E on some taxon set X is called *time-consistent*, if there exists a time-consistent labeling of its nodes by a mapping $\tau : V \to \mathbb{N}$ such that (cf. Figure 5):

- $\tau(u) < \tau(v)$ for all directed tree edges $(u, v) \in E$,
- $\tau(u) = \tau(v)$ for all reticulation edges.

Remark. The mapping $\tau : V \to \mathbb{N}$ in Definition 14 can be interpreted as a *time stamp* and 'describes a possible ordering of speciation and reticulation events, but not necessarily actual times' (Huson et al. [20], p. 167). The first condition implies that the time stamp of a tree node is larger than the time stamp of its parent node, while the second condition means that the parents of a reticulation node, and the reticulation node itself, all have the same time stamp (cf. Huson et al. [20], p. 167).
However, when considering weighted phylogenetic networks, where edge weights explicitly represent time, these conditions translate to the following:

- The sum of branch lengths from the root to a tree node v is greater than the sum of branch lengths from the root to the parent node of v.
- The sum of branch lengths from the root to a reticulation node r equals the sum of the branch lengths from the root to each of the parent nodes of r.

Networks that are both tree-child and time-consistent are called *TCTC-networks*.

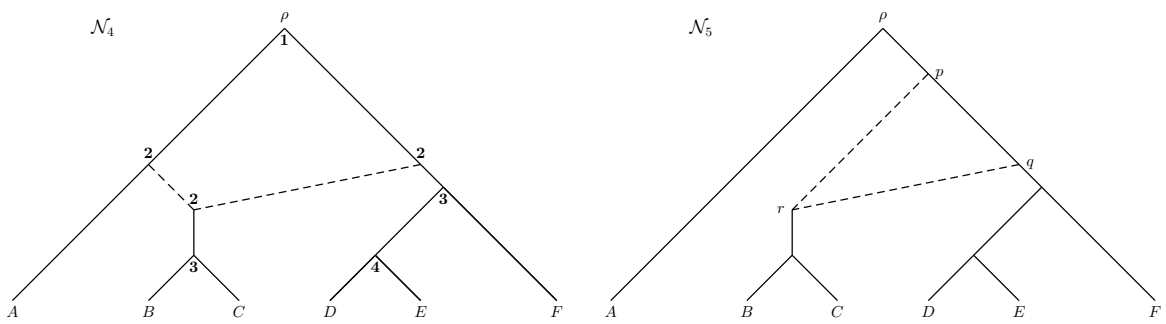


Fig. 5: Rooted phylogenetic networks $\mathcal{N}_4$ and $\mathcal{N}_5$ on $X = \{A, B, C, D, E, F\}$. All internal nodes of $\mathcal{N}_4$ can be consistently labeled by time stamps (bold arabic numerals), thus, $\mathcal{N}_4$ is a *time-consistent* network. For the network $\mathcal{N}_5$ no consistent labeling by time stamps exists. This is due to the fact that the parent p of the reticulation node r is also a parent of q, the other parent of r.

3.2. Networks and their embedded trees

Mathematically, phylogenetic networks are a generalization of phylogenetic trees. However, analyzing the treelike content of a network, is often a first step in understanding the network and will be of importance in the following sections. Therefore, we now turn our attention to phylogenetic trees that are embedded in a network.

Definition 15 (Rooted trees displayed by a rooted network)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on a taxon set X and let $\mathcal{T}$ be a phylogenetic X-tree. We say

that $\mathcal{T}$ is *embedded in* $\mathcal{N}$, or that $\mathcal{N}$ *displays* $\mathcal{T}$, if $\mathcal{T}$ can be obtained from $\mathcal{N}$ by deleting one of the reticulation edges for each reticulation node and suppressing the resulting nodes of indegree 1 and outdegree 1.
We use $\mathsf{T}(\mathcal{N})$ to denote the (multi)set of all rooted phylogenetic trees on X displayed by $\mathcal{N}$.

Remarks.

- Note that if there are k reticulation nodes in a rooted binary network $\mathcal{N}$ on a taxon set X, then there are at most 2^k phylogenetic X-trees displayed by $\mathcal{N}$. However, this bound does not have to be sharp. In fact, it is computationally hard ($\#P$-complete)[3] to calculate the number of trees embedded in a given network (cf. Linz et al. [25]).
  Furthermore, deciding whether a given binary phylogenetic X-tree $\mathcal{T}$ is displayed by a given phylogenetic network $\mathcal{N}$ on X is an NP-complete problem (cf. Kanj et al. [23]).

- There are phylogenetic networks that display a phylogenetic tree twice, i.e. there are two distinct sets of reticulation edges in the network $\mathcal{N}$ that yield the same tree when being deleted (cf. Cordue et al. [9]). Thus, $\mathsf{T}(\mathcal{N})$ is possibly a multiset (cf. Figure 6, Figure 7).

- When deleting one of the reticulation edges for each reticulation node in a network $\mathcal{N}$ on X, we may obtain a tree $\mathcal{T}$, where some internal nodes of $\mathcal{N}$ have become leaves with indegree 1 and outdegree 0. In this case, however, we do not regard $\mathcal{T}$ as an embedded X-tree, because the former internal nodes, which have become leaves, do not belong to the taxon set X, on which $\mathcal{N}$ is defined. In particular, the taxon set of an embedded X-tree $\mathcal{T}$ cannot be different from the taxon set X of $\mathcal{N}$.
  We remark that an alternative approach would be to delete all leaves that do not belong to X, together with their incident edges and regard the resulting tree as displayed by $\mathcal{N}$. However, when considering phylogenetic networks with edge weights, this method can lead to trees in which the sum of the edge lengths, i.e. the total *phylogenetic diversity*, does not equal the sum of the edge lengths in the original phylogenetic network (cf. Figure 6). Thus, using this approach, we would lose some of the information present in the network. We therefore suggest to use the first approach for weighted phylogenetic networks, i.e. we suggest that only X-trees may be regarded as displayed by the network $\mathcal{N}$, while all other

[3] $\#P$ is a complexity class associated with counting problems.

X'-trees obtained when deleting one reticulation edge for each reticulation node, with $X' \neq X$, may be discarded.

Note, however, that the problem of internal nodes becoming leaves does only affect phylogenetic networks that are not tree-child. For an internal node v to become a leaf in the process of deleting reticulation edges for each reticulation node, the node v must be the parent of two reticulation nodes. In particular, v cannot have a tree node as a child and thus, the network cannot be tree-child.

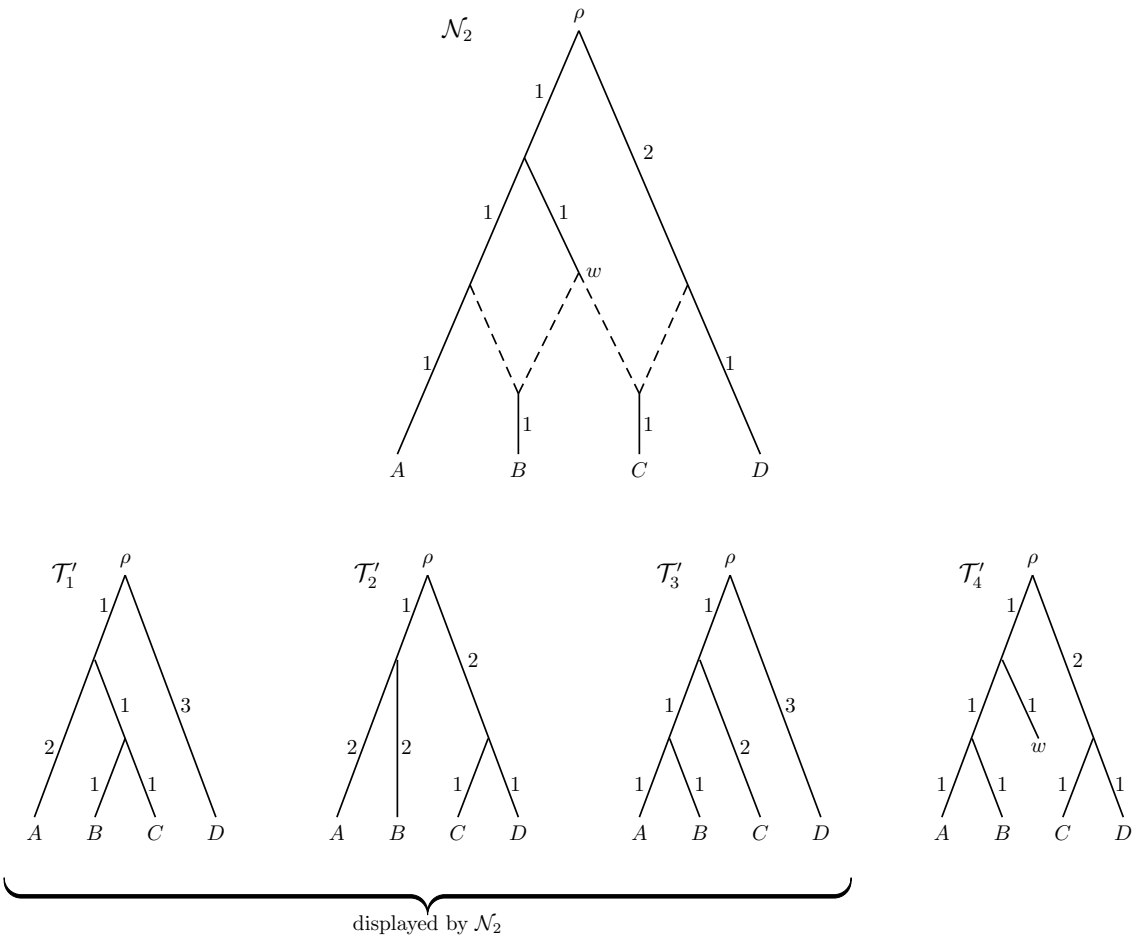


Fig. 6: The rooted phylogenetic network $\mathcal{N}_2$ on $X = \{A, B, C, D\}$ displays the phylogenetic X-trees $\mathcal{T}_1', \mathcal{T}_2'$ and $\mathcal{T}_3'$. When deleting exactly one reticulation edge for each of the two reticulation nodes, we also obtain the tree $\mathcal{T}_4'$, in which the internal node w of $\mathcal{N}$ has become a leaf. However, we do not regard $\mathcal{T}_4'$ as a phylogenetic X-tree displayed by $\mathcal{N}_2$, because w does not belong to taxon set X. Thus, in this case we have $\mathsf{T}(\mathcal{N}_2) = \{\mathcal{T}_1', \mathcal{T}_2', \mathcal{T}_3'\}$. In particular, $|\mathsf{T}(\mathcal{N}_2)| = 3 < 4 = 2^2$.
Alternatively, we could delete leaf w and its incident edge in $\mathcal{T}_4'$ and regard the resulting tree as displayed by $\mathcal{N}_2$. However, by deleting w and the edge incident to it, we retrieve a tree, where the sum of branch lengths is 8, while the sum of branch lengths for $\mathcal{N}_2$ and the trees $\mathcal{T}_1', \mathcal{T}_2'$ and $\mathcal{T}_3'$ is 9. Thus, we lose some information.
Note that if we omit branch lengths and only consider the treeshape of the displayed trees, deleting the node w and its incident edge in $\mathcal{T}_4'$ yields tree $\mathcal{T}_2'$. Thus, in this case, we would say that $\mathcal{T}_2'$ was displayed twice by $\mathcal{N}_2$. In particular, $\mathsf{T}(\mathcal{N}_2) = \{\mathcal{T}_1', \mathcal{T}_2', \mathcal{T}_2', \mathcal{T}_3'\}$ would be a multiset.

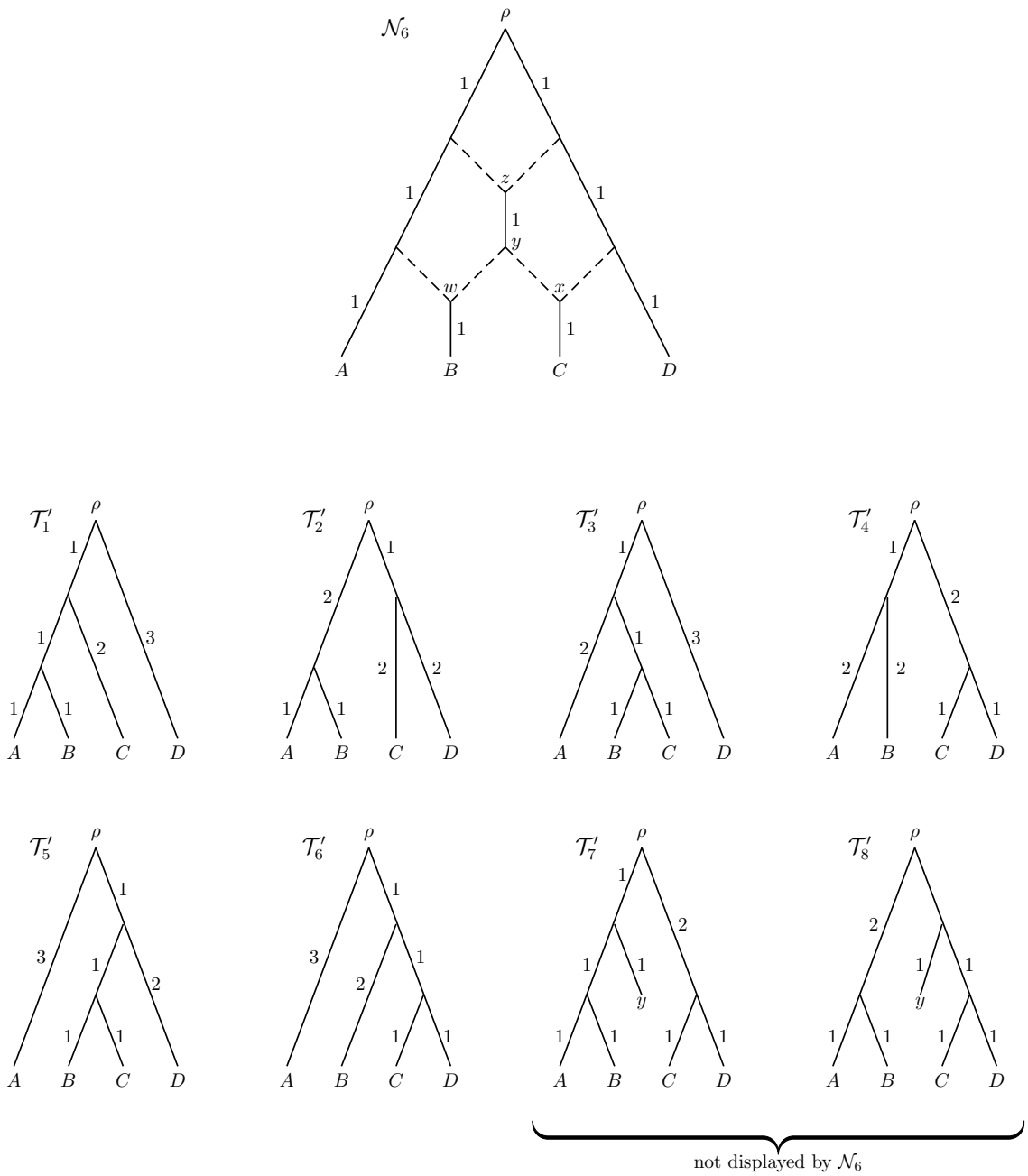


Fig. 7: The rooted phylogenetic network $\mathcal{N}_6$ on $X = \{A, B, C, D\}$ displays the phylogenetic X-trees $\mathcal{T}_1', \ldots, \mathcal{T}_6'$, while $\mathcal{T}_7'$ and $\mathcal{T}_8'$ are not displayed by $\mathcal{N}_6$. Thus, $|\mathsf{T}(\mathcal{N}_6)| = 6 \leq 8 = 2^3$. When branch lengths are omitted and only the treeshapes are considered, we have $\mathcal{T}_2' = \mathcal{T}_4'$. Thus, we would say that this treeshape was displayed twice by $\mathcal{N}_6$, because two distinct sets of reticulation edges in the network $\mathcal{N}_6$ yield this treeshape when being deleted.

3.3. The LSA tree

We now introduce one last concept, which will be used in the following, namely the *lowest stable ancestor tree*, or *LSA tree* for short.

Recall that for a rooted phylogenetic tree $\mathcal{T}$ the unique *lowest common ancestor (LCA)* of any two nodes u and v is defined as the lowest (i.e. deepest) node $lca(u, v)$ that is both an ancestor of u and v.
However, for a rooted phylogenetic network $\mathcal{N}$ the lowest common ancestor of two nodes does not have to be unique (cf. Figure 8), but is defined as the set of all lowest nodes that are ancestors of both u and v (cf. Huson et al. [20], p. 140).

In contrast to the LCA, the *lowest stable ancestor (LSA)* is uniquely determined. It is defined as follows:

Definition 16 (Lowest stable ancestor (cf. Huson et al. [20], p. 142))**.** Let $\mathcal{N}$ be a rooted phylogenetic network and let u be a node of $\mathcal{N}$ that is not the root. Then a *stable ancestor* of v is a any ancestor of v that lies on all directed paths from the root to v. The *lowest stable ancestor (LSA)* of v is defined as the last node $lsa(v)$ that is contained on all paths from the root to v, excluding v.

Remarks.

- Note that the lowest common ancestor is defined for any *pair* of nodes, while the lowest stable ancestor is defined for a *single* node.
- The lowest stable ancestor of any tree node in a phylogenetic network is its parent node.

We can now introduce the *LSA tree* as defined in Huson et al. [20] (p. 142).

Definition 17 (LSA tree)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on a taxon set X. The *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N})$ associated with $\mathcal{N}$ is a rooted phylogenetic X-tree that can be computed as follows: For each reticulation node r in $\mathcal{N}$, remove all edges directed into r and add a new edge $e = (lsa(r), r)$ from the lowest stable ancestor of r into r. Then repeatedly remove all unlabeled leaves and nodes with in- and outdegree 1, until no further such removal is possible (cf. Figure 9).

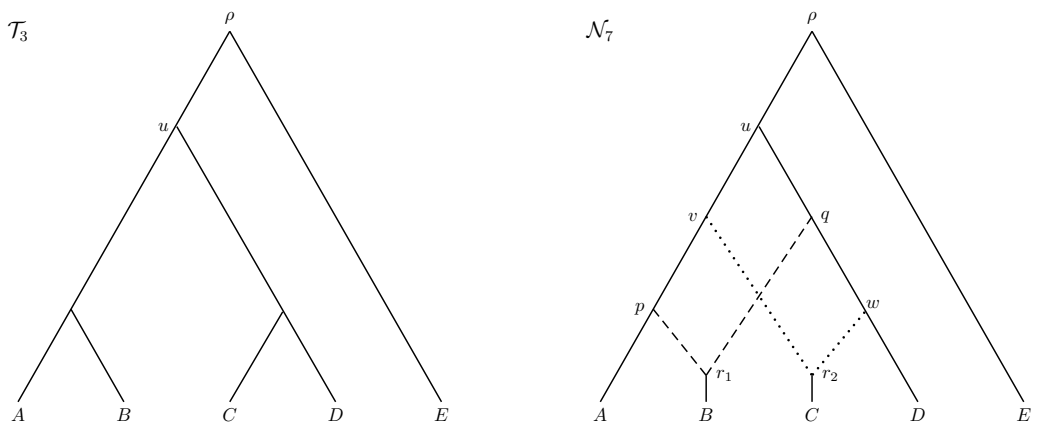


Fig. 8: Rooted phylogenetic tree $\mathcal{T}_3$ and rooted phylogenetic network $\mathcal{N}_7$ on $X = \{A, B, C, D, E\}$. The *lowest common ancestor* of B and C in $\mathcal{T}_3$ is the node u. In $\mathcal{N}_7$ the leaves B and C have two *lowest common ancestors*, namely the nodes v and q. However, their *lowest stable ancestor* is uniquely determined. The *lowest stable ancestor* of B is the node r_1 and the *lowest stable ancestor* of C is the node r_2. The node u, however, is the *lowest stable ancestor* of the reticulation nodes r_1 and r_2 (taken from Huson et al. [20], p. 141).

Remarks.

- The *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N})$ of a rooted binary phylogenetic network $\mathcal{N}$ is not necessarily binary (cf. Figure 9).
- We remark that this concept can also be used to construct a *consensus tree* for a set of rooted phylogenetic X-trees. We refer the reader to Huson et al. [20] (p. 142) for more details and just briefly comment on the idea of this method: To obtain the *LSA consensus tree*, a so-called *cluster network* for the set of input trees is constructed and then the *LSA tree* is returned. The LSA consensus tree is, however, closely related to the *Adams consensus tree* (again, cf. Huson et al. [20], p. 65, p. 143).

So far, the *LSA tree* associated with a phylogenetic network has only been considered for unweighted networks, i.e. networks that do not come with branch lengths. In order to use the *LSA tree* for the analysis of *phylogenetic diversity*, we first need to generalize this concept to weighted networks.

In order to do so, we have to assign an edge length to each new edge $e = (lsa(r), r)$ between a reticulation node r and $lsa(r)$. We suggest to set this edge length to the average length of a path between the reticulation node r and its lowest stable ancestor $lsa(r)$.

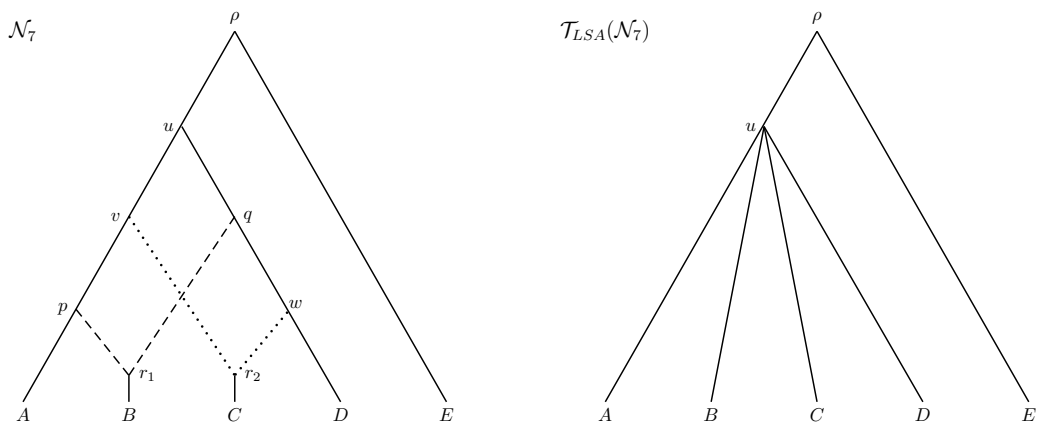


Fig. 9: Rooted phylogenetic network $\mathcal{N}_7$ on $X = \{A, B, C, D, E\}$ and its *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N}_7)$. The node u is the lowest stable ancestor of the reticulation nodes r_1 and r_2. In order to construct the *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N}_7)$, all edges directed into r_1 and r_2 are removed and two new edges $e_1 = (u, r_1)$ and $e_2 = (u, r_2)$ are added. Then the nodes v, p, q, w, r_1 and r_2, which now have in- and outdegree 1, are removed. Note that the X-tree $\mathcal{T}_{LSA}(\mathcal{N}_7)$ is not binary, as the internal node u has outdegree four.

Definition 18 (Weighted LSA Tree)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on some taxon set X. The *weighted LSA Tree* $\mathcal{T}_{LSA}(\mathcal{N})$ associated with $\mathcal{N}$ can be computed by removing all in-edges for each reticulation node r and adding a new in-edge $e = (lsa(r), r)$ from the *LSA* of r to r and then repeatedly removing all unlabeled leaves and all nodes with both in- and out-degree one. Thereby the length of the new edge e is set to the mean of path lengths of paths between $lsa(r)$ and r, i.e.

$$length(e = (lsa(r), r)) := \frac{1}{|\mathcal{P}_r|} \sum_{P \in \mathcal{P}_r} length(P), \tag{3.1}$$

where $\mathcal{P}_r$ is the set of all $lsa(r) - r-$paths P in $\mathcal{N}$. In the removal of unlabeled leaves and nodes with in- and out-degree one, formerly distinct edges may be melted into a new edge, in which case, their edge lengths are added to yield the edge length of the new edge (cf. Figure 10).

Remarks.

- If we speak of the *LSA tree* in the following, we will always mean the *weighted LSA Tree.*
- Instead of assigning the average length of a path from a reticulation node r to its lowest stable ancestor $lsa(r)$ to the edge $e = (lsa(r), r)$ in the *LSA tree*, other lengths could be used (e.g. the maximum path length, the minimum path

length etc.). However, we will not further analyze the effect of different methods in constructing a weighted *LSA tree*, but use the average path length from r to $lsa(r)$ as edge length of the edge $e = (lsa(r), r)$ in all cases.

- Constructing a weighted *LSA tree* associated with a weighted phylogenetic network $\mathcal{N}$ as described above can have the effect that the sum of branch lengths in the *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N})$ does not equal the sum of branch lengths in $\mathcal{N}$ (cf. Figure 10).

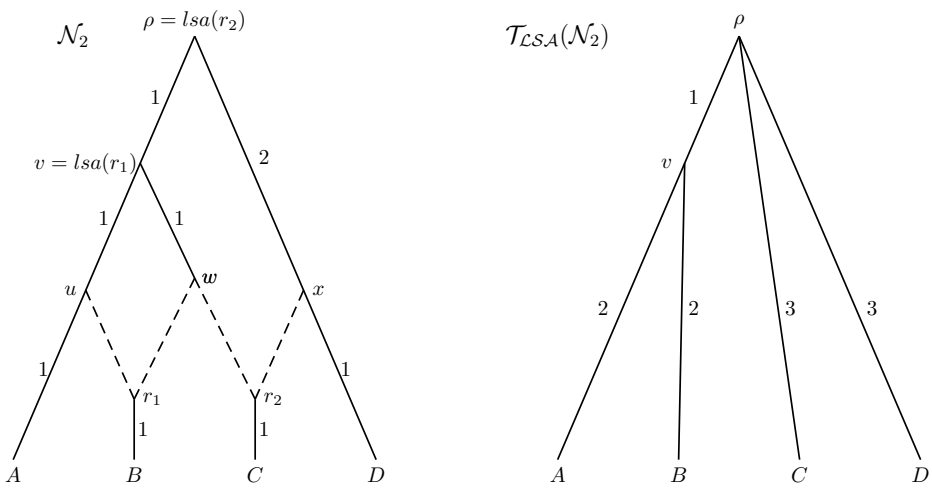


Fig. 10: Rooted binary weighted network $\mathcal{N}_2$ on $X = \{A, B, C, D\}$ and its associated weighted *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N}_2)$. The node v is the lowest stable ancestor of the reticulation node r_1 and we have to consider two paths when calculating the length of the edge $e = (lsa(r_1), r_1)$: $P_1 = ((v, u), (u, r_1))$ with $length(P_1) = 1 + 0 = 1$ (recall that we have defined the lengths of reticulation edges to be zero) and $P_2 = ((v, w), (w, r_1))$ with length $length(P_2) = 1 + 0 = 1$. Thus, taking the average, we set $length((lsa(r_1), r_1)) := 1$. Analogously, the node ρ is the lowest stable ancestor of r_2 and we have to consider the paths $P_3 = ((\rho, v)(v, w)(w, r_2))$ with $length(P_3) = 1 + 1 + 0 = 2$ and $P_4 = ((\rho, x), (x, r_2))$ with $length(P_4) = 2 + 0 = 2$. Thus, we set $length((lsa(r_2), r_2)) := 2$. However, subsequently the edges (v, r_1) and (r_1, B) are merged into a new edge (v, B) of length $1+1 = 2$ and analogously, the edges (ρ, r_2) and (r_2, C) are replaced by a new edge (ρ, C) of length $2+1 = 3$ to finally yield the *LSA tree* associated with $\mathcal{N}_2$.Note that the sum of branch lengths in $\mathcal{N}_2$ is 9, while the sum of branch lengths in $\mathcal{T}_{LSA}(\mathcal{N}_2)$ is 11. However, this effect does not only occur when using the average path length of a path from a reticulation node r to its lowest stable ancestor $lsa(r)$ as edge length for the newly created edge $(lsa(r), r)$, but also when considering, for example, the maximum or minimum path length. In case of $\mathcal{N}_2$ both the use of the minimum or the maximum path length would have led to the same tree $\mathcal{T}_{LSA}(\mathcal{N}_2)$ as our method of using the average path length.

Remark. The concept of lowest stable ancestors and the lowest stable ancestor tree is not only used in phylogenetics. On the contrary, it is an important concept in the theory of flow graphs, where the *LSA* of a node v is called the *immediate dominator* of v and the *LSA tree* is referred to as the *dominator tree*. In this context the computation of the dominator tree has been extensively studied and fast algorithms exist (cf. Lengauer and Tarjan [24], Cooper et al. [8]).

Lengauer and Tarjan [24] have, for example, developed an $O(m \log n)$ algorithm for finding dominators, where n is the number of vertices of the graph and m the number of edges.

Note, however, that computing the edge lengths of the *LSA tree* adds complexity to the problem. The calculation of the average path length between a reticulation node r and its lowest stable ancestor $lsa(r)$ involves the computation of the lengths of all paths between $lsa(r)$ and r. In the worst case, the number of paths between $lsa(r)$ and r can be exponential in the number of reticulation nodes (cf. Figures 11, 12). Thus, the construction of the *LSA tree* may be infeasible for phylogenetic networks with a high number of reticulation nodes.

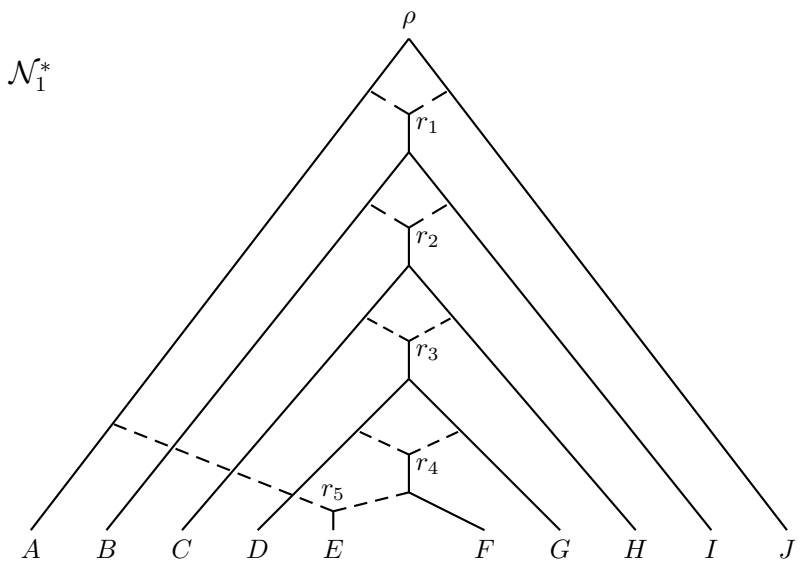


Fig. 11: Rooted binary phylogenetic network $\mathcal{N}_1^*$ on $X = \{A, B, \ldots, J\}$ with five reticulation nodes $r_1, \ldots, r_5$. The *lowest stable ancestor* of r_5 is the root ρ and there are $17 = 2^{5-1} + 1 = 2^{r(\mathcal{N}_1^*)-1} + 1$ paths from ρ to r_5. The example can be extended to an arbitrary number of reticulation nodes. In all cases, it is possible to construct a network $\mathcal{N}'$ with $2^{r(\mathcal{N}')-1} + 1$ paths between the lowest stable ancestor $lsa(r)$ of a fixed reticulation node and the reticulation node r. Note that in this example the number of paths between $lsa(r_i)$ and r_i, $i = 1, 2, 3, 4$, equals two. Thus, r_5 is the only 'worst case reticulation node'.

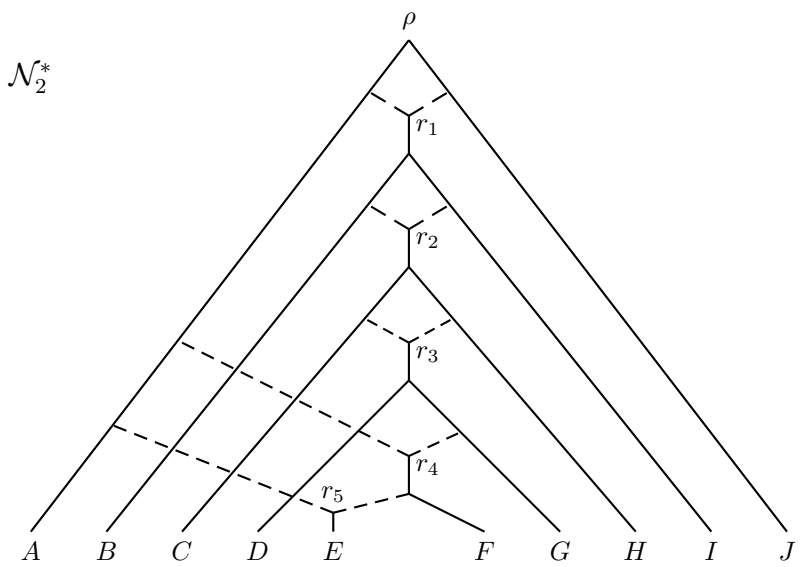


Fig. 12: Rooted binary phylogenetic network $\mathcal{N}_2^*$ on $X = \{A, B, \ldots, J\}$ with five reticulation nodes $r_1, \ldots, r_5$. The root ρ is the lowest stable ancestor of r_4 and r_5. Both for r_4 and r_5, there are $9 = 2^{r(\mathcal{N}_2^*)-2} + 1$ paths from ρ to r_4 and r_5, respectively. For r_1, r_2 and r_2, the number of paths between $lsa(r_i)$ and r_i, $i = 1, 2, 3$, equals two. Again, the example can be extended to an arbitrary number of reticulation nodes.

3.4. Hybridization probabilities

So far, we have considered the transmission of genetic material during hybridization to be equally likely for both parental species.
However, this assumption may not be justified at all times and for all organisms.
We therefore now introduce *hybridization probabilities* for phylogenetic networks.

Definition 19 (Hybridization probability)**.** Let $\mathcal{N}$ be a rooted binary phylogenetic network on a taxon set X. Let r be a reticulation node, i.e. a hybrid species, with parents p_1 and p_2. Then we use $\alpha_r \in (0,1)$ to denote the probability that the hybrid species inherits its genetic material (e.g. a nucleotide or a gene) from parent p_1 and we use $\beta_r = 1 - \alpha_r$ to denote the probability that the genetic material is inherited from parent p_2 (or vice versa). We call α_r and β_r *hybridization probabilities* and associate α_r with the reticulation edge (p_1, r) and analogously, we associate β_r with the reticulation edge (p_2, r) (or vice versa) (cf. Figure 13).

Remark. If no hybridization probabilities are given for a phylogenetic network with k reticulation nodes, we assume $\alpha_{r_i} = \beta_{r_i} = \frac{1}{2}$ for all reticulation nodes $r_i, i = 1, \ldots, k$ (cf. Figure 15).

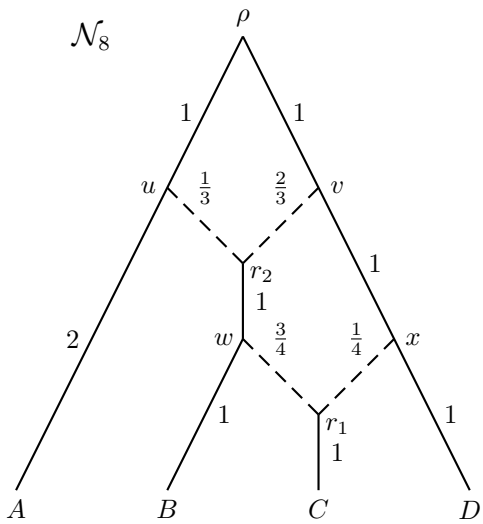


Fig. 13: Rooted binary phylogenetic network $\mathcal{N}_8$ on $X = \{A, B, C, D\}$ with *hybridization probabilities* $\alpha_{r_1} = \frac{3}{4}$ and $\alpha_{r_2} = \frac{1}{3}$. Thus, r_1 inherits its genetic material with probability $\frac{3}{4}$ from its parent w and with probability $\frac{1}{4}$ from its parent x. Analogously, r_2 inherits its genetic material with probability $\frac{1}{3}$ from its parent u and with probability $\frac{2}{3}$ from its parent v.

We now shortly define the *probability of an edge in a network*, which is clearly evident, but will be needed in the following.

Definition 20 (Probability of an edge in a network)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on a taxon set X. For each edge e (either reticulation edge or tree edge) in $\mathcal{N}$ we use $\mathbb{P}(e) \in (0,1)$ to denote the *probability of the edge* e, where

$$\mathbb{P}(e) = \begin{cases} 1, & \text{if } e \text{ is a tree edge of } \mathcal{N}; \\ \frac{1}{2}, & \text{if } e \text{ is a reticulation edge of } \mathcal{N}, \text{ but does not have a} \\ & \text{hybridization probability assigned to it;} \\ \gamma_e, & \text{if } e \text{ is a reticulation edge of } \mathcal{N} \text{ and } \gamma_e \in (0,1) \text{ is the} \\ & \text{hybridization probability assigned to } e. \end{cases}$$

We can now define the probability of an embedded tree and introduce the so-called *hybrid LSA tree* and the *Maximum Likelihood LSA tree.*

3.4.1. Probability of an embedded tree

Definition 21 (Probability of an embedded tree)**.** Let $\mathcal{N}$ be a rooted binary phylogenetic network on a taxon set X with k reticulation nodes and let $\mathsf{T}(\mathcal{N})$ be the (multi)set of phylogenetic X-trees displayed by $\mathcal{N}$. For each tree $\mathcal{T} \in \mathsf{T}(\mathcal{N})$ we re-establish the nodes of in- and outdegree 1 that were removed during its construction.[4] Then we can calculate the *probability of an embedded tree* $\mathbb{P}(\mathcal{T})$ in the following way:

1. For all $\mathcal{T} \in \mathsf{T}(\mathcal{N})$ calculate the unscaled probability

$$\mathbb{P}_{unscaled}(\mathcal{T}) = \prod_{e:\, e \text{ is contained in } \mathcal{T}} \mathbb{P}(e),$$

   where $\mathbb{P}(e)$ denotes the probability of an edge e.

2. Set $p := \sum_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} \mathbb{P}_{unscaled}(\mathcal{T})$ (scaling factor).

3. Calculate the probability

$$\mathbb{P}(\mathcal{T}) = \frac{1}{p} \cdot \mathbb{P}_{unscaled}(\mathcal{T}).$$

[4] The re-establishment of nodes of in- and outdegree 1 is necessary to trace back the edges of $\mathcal{N}$ present in $\mathcal{T} \in \mathsf{T}(\mathcal{N})$. In practice, it suffices to know, which of the reticulation edges of $\mathcal{N}$ were kept in the construction of $\mathcal{T}$, as they determine the probability of $\mathcal{T}$ (cf. Figures 14, 15).

Remark. The scaling factor $p = \sum_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} \mathbb{P}_{unscaled}(\mathcal{T})$ in Definition 21 ensures that the probabilities of all embedded trees sum up to 1, because

$$\begin{aligned}
\sum_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} \mathbb{P}(\mathcal{T}) &= \sum_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} \big(\frac{1}{p} \cdot \mathbb{P}_{unscaled}(\mathcal{T})\big) \\
&= \frac{1}{p} \sum_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} \mathbb{P}_{unscaled}(\mathcal{T}) \\
&= \frac{1}{\sum_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} \mathbb{P}_{unscaled}(\mathcal{T})} \sum_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} \mathbb{P}_{unscaled}(\mathcal{T}) \\
&= 1.
\end{aligned}$$

This is of importance if $|\,\mathsf{T}(\mathcal{N})\,| < 2^k$, i.e. if removing one reticulation edge for each reticulation node and suppressing nodes of both indegree 1 and outdegree 1 results in one or several trees, which are not phylogenetic X-trees (cf. Figure 15).
On the other hand, if $|\,\mathsf{T}(\mathcal{N})\,| = 2^k$, we have $\sum_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} \mathbb{P}(\mathcal{T}) = 1$ and thus, the scaling factor p equals 1. In this case the probability of an embedded tree $\mathcal{T} \in \mathsf{T}(\mathcal{N})$ can directly be calculated as

$$\mathbb{P}(\mathcal{T}) = \prod_{e:\, e \text{ is contained in } \mathcal{T}} \mathbb{P}(e),$$

thus, in this case we can calculate the probability of an embedded tree $\mathcal{T} \in \mathsf{T}(\mathcal{N})$ by multiplying the probabilities of its edges (cf. Figure 14).

Definition 22 (Most likely embedded tree)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on a taxon set X and let $\mathsf{T}(\mathcal{N})$ be the (multi)set of all rooted phylogenetic X-trees displayed by $\mathcal{N}$. For all $\mathcal{T} \in \mathsf{T}(\mathcal{N})$ let $\mathbb{P}(\mathcal{T})$ denote the probability of $\mathcal{T}$. Then we call

$$\mathcal{T}^* = \operatorname*{argmax}_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} \mathbb{P}(T)$$

the *most likely embedded tree* (cf. Figure 14). If the argmax is not unique, we arbitrarily choose one of the embedded trees with maximum probability.

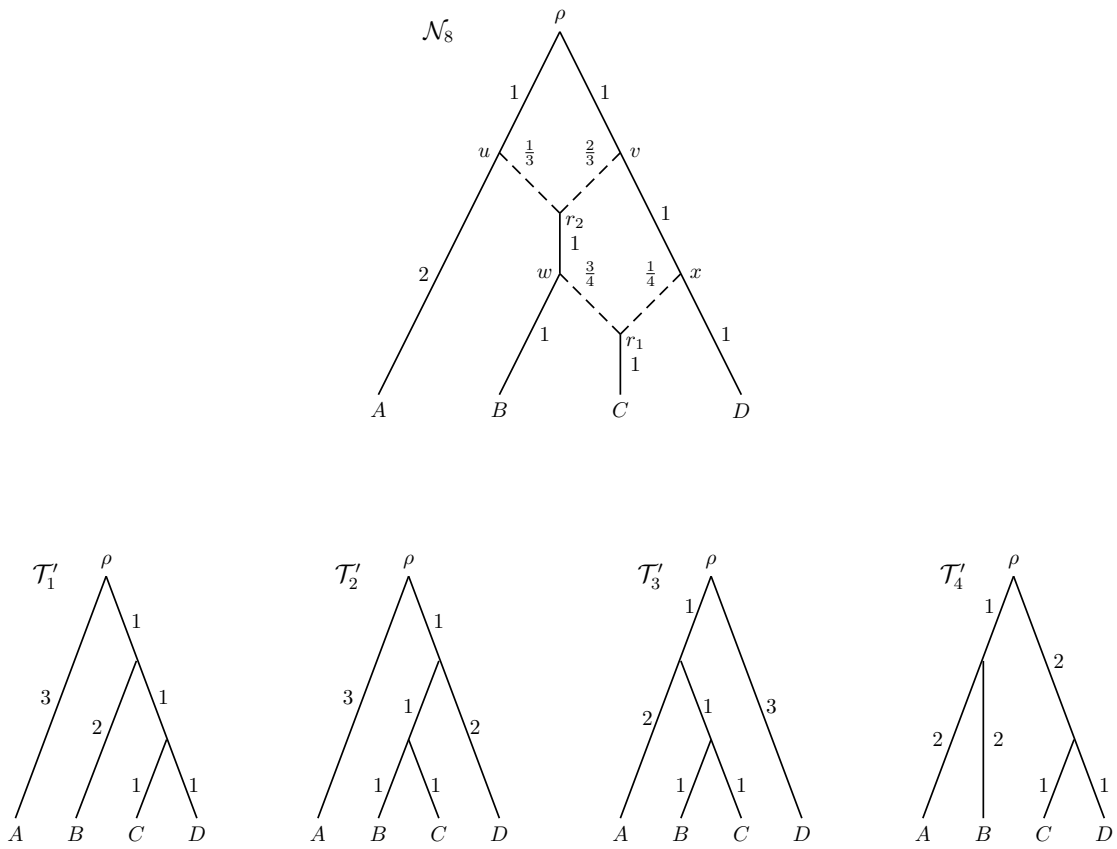


Fig. 14: The rooted binary phylogenetic network $\mathcal{N}_8$ on $X = \{A, B, C, D\}$ with two reticulation nodes displays $4 = 2^2$ phylogenetic X-trees. Applying Definition 21 we retrieve the following *probabilities of the embedded trees*: $\mathbb{P}(\mathcal{T}_1') = \frac{2}{3} \cdot \frac{1}{4} = \frac{1}{6}$, $\mathbb{P}(\mathcal{T}_2') = \frac{2}{3} \cdot \frac{3}{4} = \frac{1}{2}$, $\mathbb{P}(\mathcal{T}_3') = \frac{1}{3} \cdot \frac{3}{4} = \frac{1}{4}$ and $\mathbb{P}(\mathcal{T}_4') = \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{12}$. Note, that the probabilities of the embedded trees sum up to 1, because $\sum_{\mathcal{T} \in \mathsf{T}(\mathcal{N}_8)} \mathbb{P}(\mathcal{T}) = \mathbb{P}(\mathcal{T}_1') + \mathbb{P}(\mathcal{T}_2') + \mathbb{P}(\mathcal{T}_3') + \mathbb{P}(\mathcal{T}_4') = \frac{1}{6} + \frac{1}{2} + \frac{1}{4} + \frac{1}{12} = 1$.
$\mathcal{T}_2'$ is the *most likely embedded tree*.

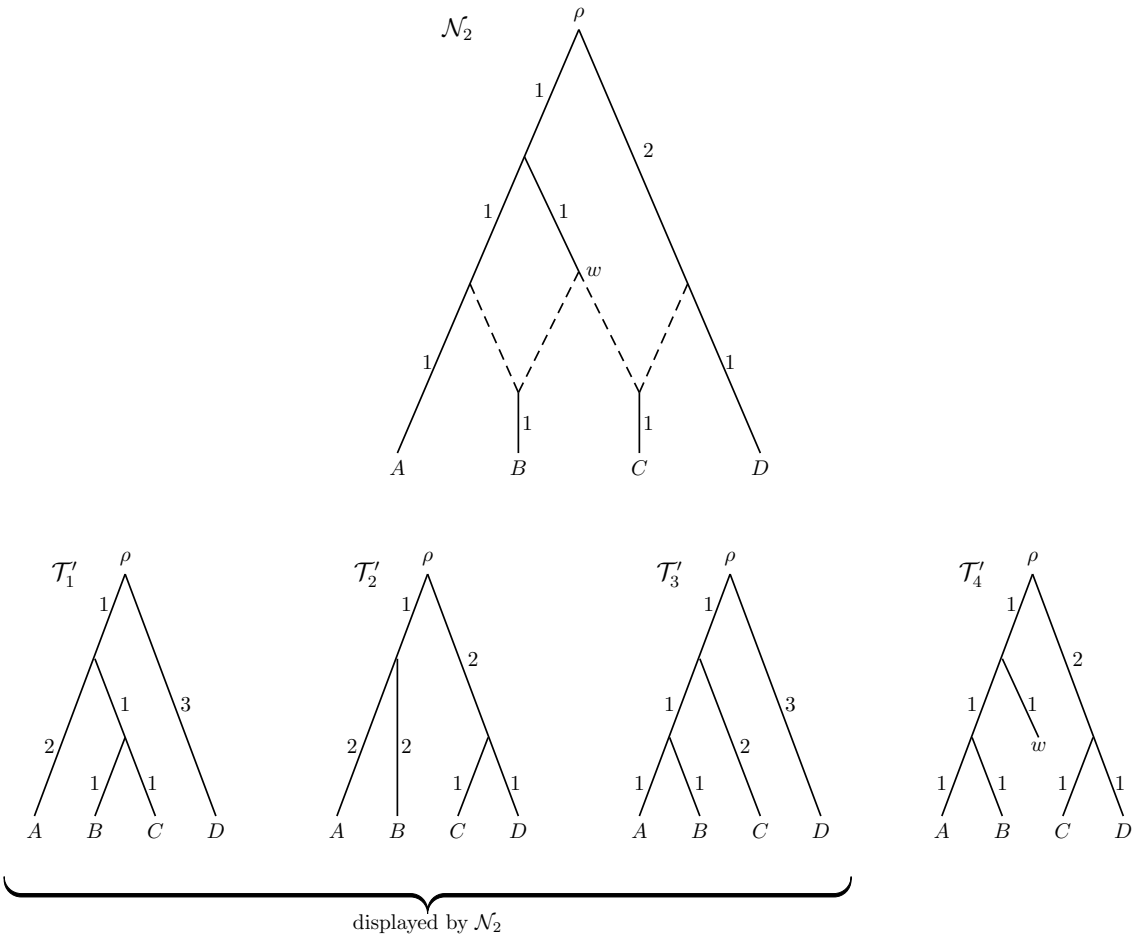


Fig. 15: The rooted binary phylogenetic network $\mathcal{N}_8$ on $X = \{A, B, C, D\}$ with two reticulation nodes displays $3 < 4 = 2^2$ phylogenetic X-trees, thus, we have to scale the probabilities of the embedded trees. As $\mathcal{N}_2$ has no hybridization probabilities assigned to its reticulation edges, we assume $\gamma_e = \frac{1}{2}$ for all reticulation edges $e \in E(\mathcal{N}_2)$ (cf. Remark after Definition 19). We have $\mathbb{P}_{unscaled}(\mathcal{T}_i') = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$, $i \in \{1, 2, 3\}$, and the scaling factor p calculates as $p := \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}$. Thus, the *probabilities of the embedded trees* are $\mathbb{P}(\mathcal{T}_i') = \frac{1}{p} \cdot \frac{1}{4} = \frac{4}{3} \cdot \frac{1}{4} = \frac{1}{3}$, $i \in \{1, 2, 3\}$. Note, that the probabilities of the embedded trees sum up to 1, because $\sum_{\mathcal{T} \in \mathsf{T}(\mathcal{N}_2)} \mathbb{P}(\mathcal{T}) = \mathbb{P}(\mathcal{T}_1') + \mathbb{P}(\mathcal{T}_2') + \mathbb{P}(\mathcal{T}_3') = \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1$.

Remark. Under certain topological constraints on the network $\mathcal{N}$, to be precise if $\mathcal{N}$ is a tree-child network (cf. Definition 12), we can directly compute a most likely embedded tree $\mathcal{T}' \in \mathsf{T}(\mathcal{N})$ (not necessarily unique), without having to consider the probabilities of all embedded trees.

Recall that the probability of an embedded tree $\mathcal{T} \in \mathsf{T}(\mathcal{N})$ was calculated by multiplying the probabilities of its edges and scaling the resulting probability, if necessary (cf. Definition 21). Essentially, it suffices to calculate the product of probabilities of the reticulation edges kept in the construction of $\mathcal{T}$, because the probability of any tree edge equals 1 and thus, the tree edges do not determine the probability of an embedded tree.

Thus, a greedy approach may be used to construct a most likely embedded tree $\mathcal{T}' \in \mathsf{T}(\mathcal{N})$, where for each reticulation node r the reticulation edge with the highest associated hybridization probability is kept, while the other edge is discarded (if both reticulation edges directed into r have the same probability, we arbitrarily choose one to keep). As a result, the product of the probabilities of the reticulation edges will be maximal, and thus, $\mathcal{T}'$ will be a most likely embedded tree.

Note that this approach only works, if the network under consideration is a tree-child network. If $\mathcal{N}$ is not a tree-child network, the greedy approach of keeping the most likely reticulation edge for each reticulation node r and discarding the other edge, may result in a tree that is not a phylogenetic X-tree, and thus, not displayed by $\mathcal{N}$ (cf. Figure 16).

However, if $\mathcal{N}$ is a tree-child network, this problem does not occur (cf. Remark on page 20).

Example 7. Consider the tree-child network $\mathcal{N}_8$ on $X = \{A, B, C, D\}$ and its embedded trees depicted in Figure 14. We have $\mathbb{P}(\mathcal{T}_1') = \frac{1}{6}$, $\mathbb{P}(\mathcal{T}_2') = \frac{1}{2}$, $\mathbb{P}(\mathcal{T}_3') = \frac{1}{4}$ and $\mathbb{P}(\mathcal{T}_4') = \frac{1}{12}$. Thus, $\mathcal{T}_2'$ is the most likely embedded tree. Note that $\mathcal{T}_2'$ contains the most likely reticulation edges for each of the two reticulation nodes, thus, we could have directly constructed $\mathcal{T}_2'$ following the greedy approach suggested above.

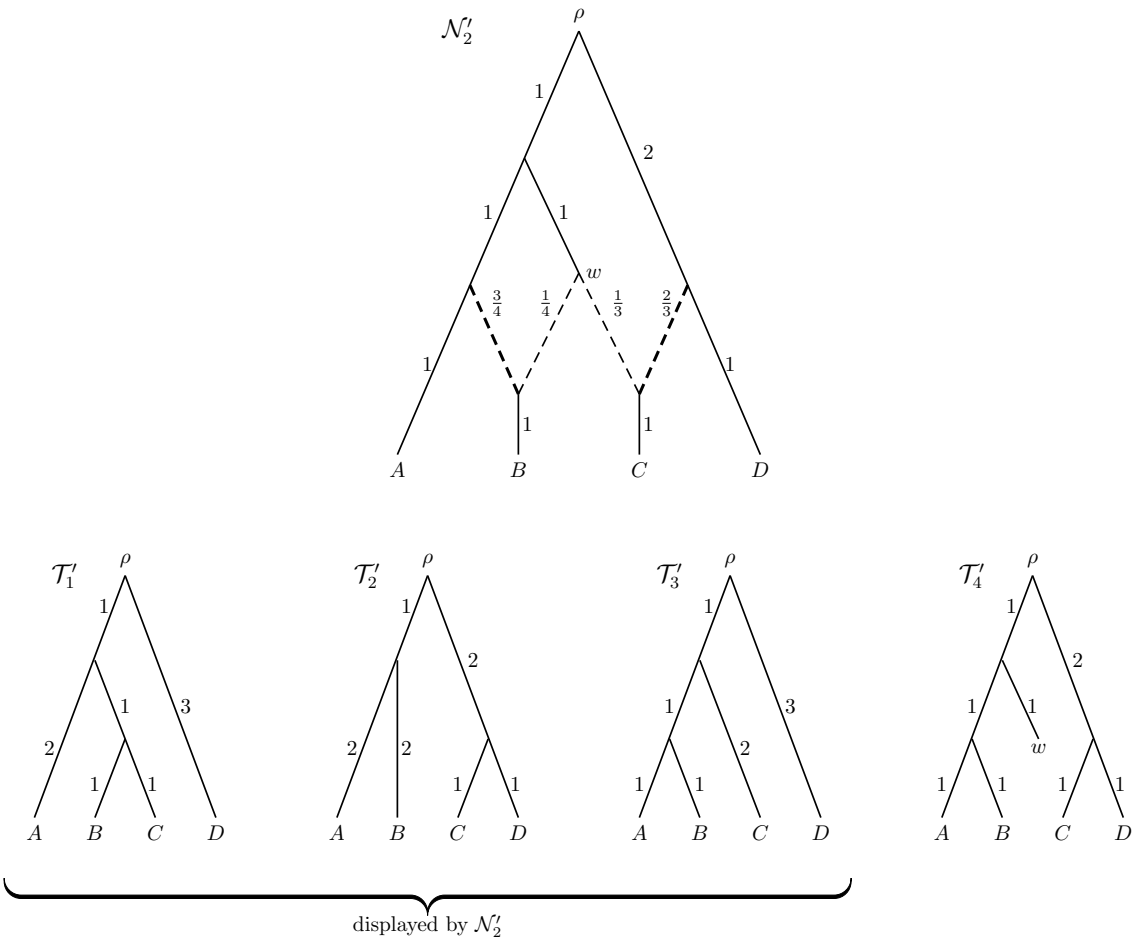


Fig. 16: Rooted phylogenetic network $\mathcal{N}_2'$ on $X = \{A, B, C, D\}$ (not tree-child) and its embedded trees $\mathcal{T}_1', \mathcal{T}_2'$ and $\mathcal{T}_3'$. We retrieve the unscaled probabilities $\mathbb{P}_{unscaled}(\mathcal{T}_1') = \frac{1}{12}$, $\mathbb{P}_{unscaled}(\mathcal{T}_2') = \frac{1}{6}$ and $\mathbb{P}_{unscaled}(\mathcal{T}_3') = \frac{1}{4}$ and thus, the scaled probabilities $\mathbb{P}(\mathcal{T}_1') = \frac{1}{6}$, $\mathbb{P}(\mathcal{T}_2') = \frac{1}{3}$ and $\mathbb{P}(\mathcal{T}_3') = \frac{1}{2}$. Thus, $\mathcal{T}_3'$ is the *most likely embedded tree.*

Following a greedy approach (cf. page 35) to construct a most likely embedded tree, and thus, keeping the most likely reticulation edge for each reticulation node (depicted as dashed bold lines), however, results in the tree $\mathcal{T}_4'$ that is not a phylogenetic X-tree.

3.4.2. Hybrid LSA tree and Maximum Likelihood LSA tree

Recall that the *LSA tree* associated with a network $\mathcal{N}$ was constructed by removing all in-edges for each reticulation node r and adding a new in-edge $e = (lsa(r), r)$ from the *LSA* of r to r and then repeatedly removing all unlabeled leaves and all nodes with both in- and out-degree one (cf. Definition 17).

Furthermore, the edge length of a new edge $e = (lsa(r), r)$ between a reticulation node r and its lowest stable ancestor $lsa(r)$ was defined as the average length of a path from $lsa(r)$ to r (cf. Equation (3.1)). Thus, we have used the *unweighted mean* of lengths of paths between $lsa(r)$ and r to define the length of the new edge $e = (lsa(r), r)$ in the *LSA tree*.

Given hybridization probabilities for a network $\mathcal{N}$, we now suggest two alternative approaches towards assigning a length to the new edge $e = (lsa(r), r)$. On the one hand, we suggest to use the *weighted mean* of path lengths between $lsa(r)$ and r, where each $lsa(r) - r-$path P is weighted according to its probability. On the other hand, we suggest to use the most likely path length between a reticulation node and its lowest stable ancestor, i.e. the length of the most likely path between them. Therefore, we now define the *probability of a path in* $\mathcal{N}$.

Definition 23 (Probability of a path in a network)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on a taxon set X. Now let u and v be two nodes in $\mathcal{N}$, where u is a ancestor of v (i.e. there exists at least one directed path from u to v). Then we define the *probability of a* $u - v-$*path* Puv *in* $\mathcal{N}$ as

$$\mathbb{P}(P_{uv}) = \prod_{\substack{e: \\ e \text{ is edge on } P_{uv}}} \mathbb{P}(e), \tag{3.2}$$

thus, we retrieve the probability of a path by multiplying the probabilities of its edges (cf. Figure 17).

Definition 24 (Hybrid LSA tree)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on a taxon set X. The *hybrid LSA tree* $\mathcal{T}_{LSA}^{hyb}(\mathcal{N})$ can be computed by removing all in-edges for each reticulation node r and adding a new in-edge $e = (lsa(r), r)$ from the *LSA* of r to r and then repeatedly removing all unlabeled leaves and all nodes with both in- and out-degree one. Thereby the length of the new edge e is set to the weighted mean of path lengths of paths between $lsa(r)$ and r, i.e.

$$length(e = (lsa(r), r)) := \sum_{P \in \mathcal{P}_r} \mathbb{P}(P) \cdot length(P), \tag{3.3}$$

where $\mathcal{P}_r$ is the set of all $lsa(r) - r-$paths and $\mathbb{P}(P)$ is the probability of any such

path. In the removal of unlabeled leaves and nodes with in- and out-degree one, formerly distinct edges may be melted into a new edge, in which case, their edge lengths are added to yield the edge length of the new edge (cf. Figures 17, 18).

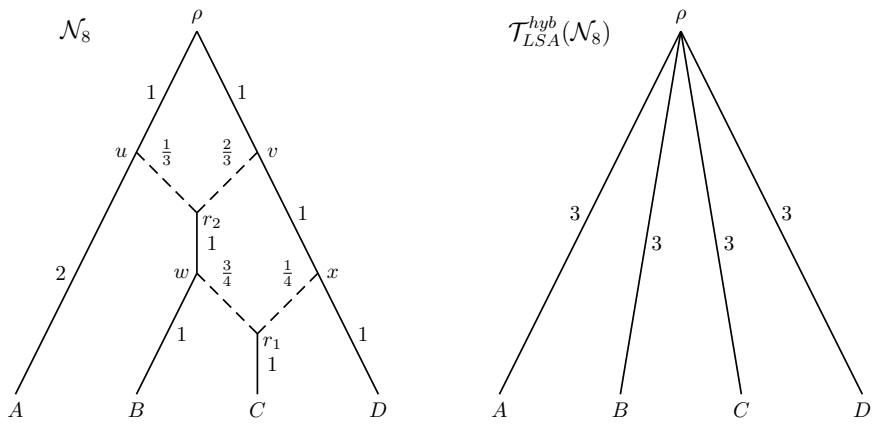


Fig. 17: Rooted binary network $\mathcal{N}_8$ on $X = \{A, B, C, D\}$ and its associated *hybrid LSA tree* $\mathcal{T}_{LSA}^{hyb}(\mathcal{N}_8)$. The root ρ is the lowest stable ancestor of the reticulation node r_1 and we have to consider three paths when calculating the length of the new edge $e = (\rho, r_1)$: $P_1 = \big((\rho, u), (u, r_2), (r_2, w), (w, r_1)\big)$ with $length(P_1) = 2$ and probability $\mathbb{P}(P_1) = 1 \cdot \frac{1}{3} \cdot 1 \cdot \frac{3}{4} = \frac{1}{4}$, $P_2 = \big((\rho, v), (v, r_2), (r_2, w), (w, r_1)\big)$ with length $length(P_2) = 2$ and probability $\mathbb{P}(P_2) = 1 \cdot \frac{2}{3} \cdot 1 \cdot \frac{3}{4} = \frac{1}{2}$ and $P_3 = \big((\rho, v), (v, x), (x, r_1)\big)$ with $length(P_3) = 2$ and probability $\mathbb{P}(P_3) = 1 \cdot 1 \cdot \frac{1}{4} = \frac{1}{4}$. Thus, we have $length(e = (\rho, r_1)) = \frac{3}{4} \cdot 2 + \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 2 = 2$. Analogously, ρ is the lowest stable ancestor of the reticulation node r_2 and similar calculations yield $length(f = (\rho, r_2)) = 1$. Removing all unlabeled leaves and nodes with in- and out-degree one and adding edge lengths of merged edges, yields the *hybrid LSA tree* $\mathcal{T}_{LSA}^{hyb}(\mathcal{N}_8)$.
Note that for $\mathcal{N}_8$ the *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N}_8)$ and the *hybrid LSA tree* $\mathcal{T}_{LSA}^{hyb}(\mathcal{N}_8)$ coincide, even though, the paths between a reticulation node and its lowest stable ancestor have different probabilities. However, as all paths between ρ and r_1 have length 2, the length of the edge $e' = (\rho, r_1)$ in the ordinary *LSA tree* would be calculated as $length(e') = \frac{1}{3}(2 + 2 + 2) = 2$ and analogously, the length of $f' = (\rho, r_2)$ in the ordinary *LSA tree* would be calculated as $length(f') = \frac{1}{2}(1 + 1) = 1$. Thus, the tree $\mathcal{T}_{LSA}(\mathcal{N}_8)$ coincides with $\mathcal{T}_{LSA}^{hyb}(\mathcal{N}_8)$. For the same reasons, the *Maximum Likelihood LSA tree* coincides with the ordinary *LSA tree* and the *hybrid LSA tree*.

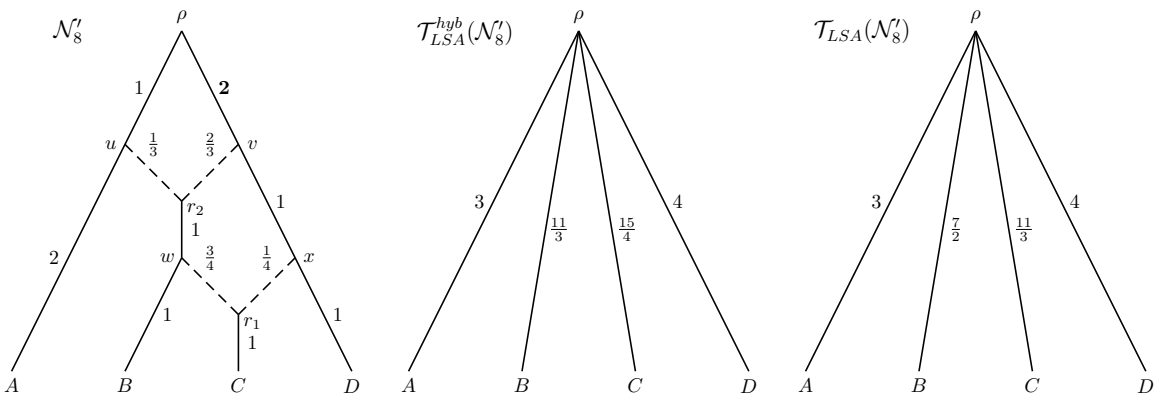


Fig. 18: Rooted binary network $\mathcal{N}_8'$ on $X = \{A, B, C, D\}$ and its associated *hybrid LSA tree* $\mathcal{T}_{LSA}^{hyb}(\mathcal{N}_8')$ and LSA tree $\mathcal{T}_{LSA}(\mathcal{N}_8')$.
Note that $\mathcal{N}_8'$ is very similar to $\mathcal{N}_8$ (cf. Figure 17), the only difference being the length of the edge (ρ, v), which is now 2 as opposed to 1 in $\mathcal{N}_8$. Again, ρ is the lowest stable ancestor of r_1 and we have to consider the three paths P_1, P_2 and P_3 (cf. Figure 17) when calculating the length of the new edge $e = (\rho, r_1)$. Thus, $length(e) = \underbrace{\frac{1}{4} \cdot 2}_{P_1} + \underbrace{\frac{1}{2} \cdot 3}_{P_2} + \underbrace{\frac{1}{4} \cdot 3}_{P_3} = \frac{11}{4}$.
For the edge $f = (\rho, r_2)$ we have $length(f) = \frac{5}{3}$. Removing all unlabeled leaves and nodes with in- and out-degree one and adding edge lengths of merged edges, yields the *hybrid LSA tree* $\mathcal{T}_{LSA}^{hyb}(\mathcal{N}_8')$.
For $\mathcal{N}_8'$ the *hybrid LSA tree* and the ordinary *LSA tree* do not coincide. When computing the ordinary *LSA tree*, we have $length(e' = (\rho, r_1)) = \frac{1}{3}(2+3+3) = \frac{8}{3}$ and $length(f' = (\rho, r_2)) = \frac{1}{2}(1+2) = \frac{3}{2}$. Thus, the resulting edge lengths in the ordinary *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N}_8')$ are different from the edge lengths in the *hybrid LSA tree* $\mathcal{T}_{LSA}^{hyb}(\mathcal{N}_8')$.

Definition 25 (Maximum Likelihood LSA tree)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on a taxon set X. The *Maximum Likelihood LSA tree* $\mathcal{T}_{LSA}^{ML}(\mathcal{N})$ can be computed by removing all in-edges for each reticulation node r and adding a new in-edge $e = (lsa(r), r)$ from the LSA of r to r and then repeatedly removing all unlabeled leaves and all nodes with both in- and out-degree one. Thereby the length of the new edge e is set to the to the length of the most likely path P^* between $lsa(r)$ and r, i.e.

$$length(e) = length(P^*), \tag{3.4}$$

with

$$P^* = \operatorname*{argmax}_{P \in \mathcal{P}_r} \mathbb{P}(P),$$

where $\mathcal{P}_r$ is the set of all $lsa(r) - r-$paths and $\mathbb{P}(P)$ is the probability of any such path. If the argmax is not unique, we choose one of the most likely $lsa(r) - r-$paths in $\mathcal{P}_r$ with minimum weight.[5] In the removal of unlabeled leaves and nodes with in- and out-degree one, formerly distinct edges may be melted into a new edge, in which case, their edge lengths are added to yield the edge length of the new edge (cf. Figure 19).

[5] Alternatively, we could arbitrarily choose one of the most likely $lsa(r) - r-$paths. However, choosing the path with minimum weight makes the results reproducible.

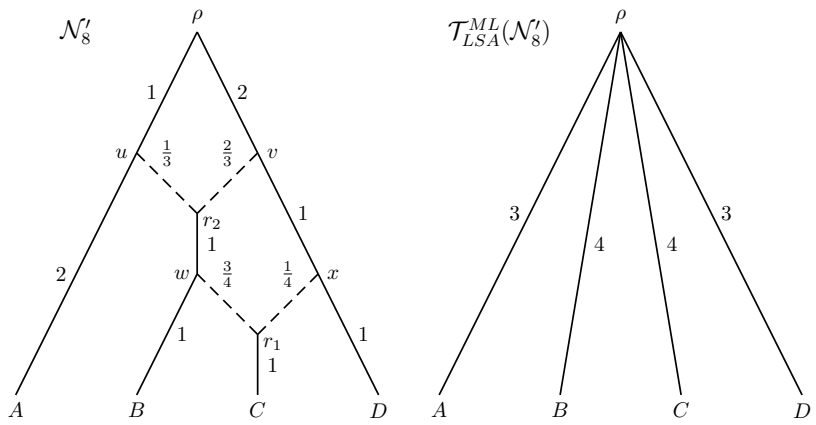


Fig. 19: Rooted binary network $\mathcal{N}_8'$ on $X = \{A, B, C, D\}$ and its associated *Maximum Likelihood LSA tree* $\mathcal{T}_{LSA}^{hyb}(\mathcal{N}_8')$ and *LSA tree* $\mathcal{T}_{ML}(\mathcal{N}_8')$.
As in Figure 17 and in Figure 18, we have to consider the three paths P_1, P_2 and P_3 when calculating the length of the new edge $e = (\rho, r_1)$. P_1 has length 2 and probability $\mathbb{P}(P_1) = \frac{1}{4}$, P_2 has length 3 and probability $\mathbb{P}(P_2) = \frac{1}{2}$ and P_3 has length 3 and probability $\mathbb{P}(P_3) = \frac{1}{4}$ (cf. Figure 17). Thus, P_2 is the most likely path between ρ and r_1 and we set $length(e = (\rho, r_1)) = 3$. Analogously, we retrieve $length(f = (\rho, r_2)) = 2$. Removing all unlabeled leaves and nodes with in- and out-degree one and adding edge lengths of merged edges, yields the *Maximum Likelihood LSA tree* $\mathcal{T}_{LSA}^{ML}(\mathcal{N}_8')$.

4. Generalization of phylogenetic diversity to hybridization networks

We are now in the position to propose different ways of generalizing the concept of *phylogenetic diversity* from trees to networks.

In a first step we directly apply the definition of *phylogenetic diversity* to networks by calculating minimum cost arborescences. Secondly, we consider the set of embedded trees to define the *phylogenetic diversity* present in a network, before using the *LSA tree* induced by a network to quantify its diversity. In all cases we consider both networks without hybridization probabilities and networks with hybridization probabilities. In case of the latter we suggest a fourth approach towards the generalization of *phylogenetic diversity* from trees to networks, namely the so-called *inherited phylogenetic diversity*.

4.1. Phylogenetic net diversity

Recall that for a rooted phylogenetic X-tree $\mathcal{T}$, the *phylogenetic diversity* of a subset $S \subseteq X$ of leaves was defined as the sum of branch lengths in the smallest spanning

tree, or, to be more precise, in the smallest arborescence (cf. Definition 3) connecting those leaves and the root.

In this case, the term *smallest arborescence* means that the arborescence does not contain any additional nodes or edges than those connecting the set S and the root.

Note, however, that a rooted phylogenetic X-tree $\mathcal{T}$ itself is an arborescence and thus, any arborescence connecting a subset $S \subseteq X$ and the root is uniquely determined. It is the sub-arborescence of $\mathcal{T}$ containing the taxa in S and the root.

This implies that the smallest arborescence connecting S and the root will automatically be a *minimum cost arborescence* in the graph-theoretical sense, i.e. an arborescence whose weight (the sum of its branch lengths) is no larger than the weight of any other arborescence connecting S and the root.

However, in case of a rooted phylogenetic network $\mathcal{N}$ on a taxon set X, there may be more than one arborescence connecting the leaves of some set $S \subseteq X$ and the root.

We therefore suggest to alter the definition of *phylogenetic diversity* for networks by replacing the term *smallest arborescence* by *minimum cost arborescence.*

Definition 26 (Phylogenetic net diversity (PND))**.** Let $\mathcal{N}$ be a rooted phylogenetic network on some taxon set X. For a subset $S \subseteq X$ of taxa we define the *phylogenetic net diversity* $PND(S)$ of S as the sum of branch lengths in the minimum cost arborescence containing S and the root.[6]

Example 8. Consider the rooted phylogenetic network $\mathcal{N}_2$ depicted in Figure 20. Exemplarily, we set $S = \{A, B\}$ and retrieve the following value for the *phylogenetic net diversity* of S:

$$PND_{\mathcal{N}_2}(\{A, B\}) = 4.$$

Note that $PND_{\mathcal{N}_2}(\{A, B, C, D\} = 8$, while the sum of branch lengths in $\mathcal{N}_2$ equals 9. Thus, the *phylogenetic net diversity* of all taxa in X does not equal the sum of branch lengths in $\mathcal{N}_2$, which may be regarded counterintuitive.

Remark. Determining the minimum cost arborescence containing a set $S \subseteq X$ of taxa and the root is formally an instance of the so-called *directed Steiner tree problem*[7]:

Definition 27 (Directed Steiner tree problem)**.** Given a directed weighted graph $G = (V, E)$, a specified root node ρ and a subset of nodes $U \subseteq V$ (called *terminals*), the objective of the *directed Steiner tree problem* is to find the minimum cost arborescence rooted at ρ and spanning all the nodes in U.

[6] A similar approach has also been suggested in cooperative game theory, where the minimum cost arborescence is used to define the cost of a coalition in a directed acyclic graph game (cf. Sziklai et al. [31]).

[7] Also referred to as the Steiner arborescence problem.

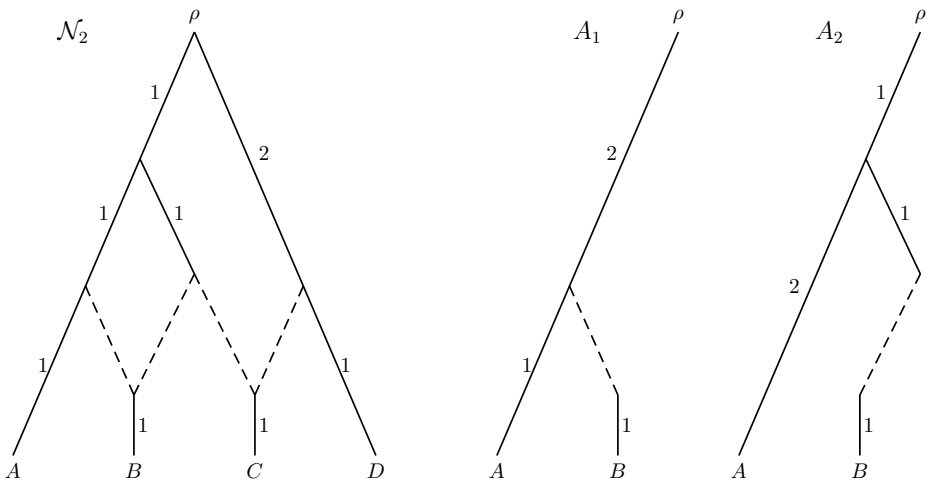


Fig. 20: Rooted phylogenetic network $\mathcal{N}_2$ on $X = \{A, B, C, D\}$ and arborescences A_1 and A_2 containing $S = \{A, B\}$ and the root. The arborescence A_1 has weight $1 + 1 + 2 = 4$, while the arborescence A_2 has weight $2 + 1 + 1 + 1 = 5$. Thus, A_1 is the *minimum cost arborescence* containing $S = \{A, B\}$ and the root ρ.

In general, the computation of a minimum Steiner arborescence is an NP-hard problem (Floudas and Pardalos [13], p. 3731) and thus, the calculation of *phylogenetic net diversity* may be infeasible. Note, however, that the calculation of the *phylogenetic diversity* of a subset $S \subseteq X$ of taxa for a phylogenetic tree $\mathcal{T}$ also involves the computation of a minimum Steiner arborescence. However, as indicated above, in this case the minimum Steiner arborescence is simply a sub-arborescence of $\mathcal{T}$ containing the taxa in S and the root. Thus, for phylogenetic trees the problem of computing a minimum Steiner arborescence spanning a set $S \subseteq X$ of taxa and the root reduces to the problem of computing a sub-arborescence containing S and the root, which can be achieved in linear time (cf. Hwang et al. [21]).

Hybrid phylogenetic net diversity

Given hybridization probabilities, we will now define the *probability of an arborescence* and then introduce the *hybrid phylogenetic net diversity*, which considers the average weight of arborescences spanning the taxa in S and the root instead of the weight of the minimum cost arborescence.

Definition 28 (Probability of an arborescence in a network)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on a taxon set X. Let $S \subseteq X$ be a subset of taxa and let A be an arborescence containing S and the root (not necessarily a minimum cost arborescence).

Then we define the *probability of the arborescence A in $\mathcal{N}$* as

$$\mathbb{P}(A) = \prod_{e:\, e \text{ is contained in } A} \mathbb{P}(e), \tag{4.1}$$

where $\mathbb{P}(e)$ is the probability of the edge e. Thus, we retrieve the probability of an arborescence by multiplying the probabilities of its edges (cf. Figure 21).

Definition 29 (Hybrid phylogenetic net diversity (PND^{hyb}))**.** Let $\mathcal{N}$ be a rooted phylogenetic network on a taxon set X. Let $S \subseteq X$ be a subset of taxa and let $\mathcal{A}_S$ be the set of all smallest arborescences spanning the taxa in S and the root, i.e. all arborescences (not necessarily of minimum cost) that span S and the root, but do not contain any additional nodes or edges than those necessary to connect the taxa in S and the root. Then we define the *hybrid phylogenetic net diversity* $PND^{hyb}(S)$ of a subset $S \subseteq X$ of taxa as the weighted mean of weights of the arborescences in $\mathcal{A}_S$, i.e.

$$PND^{hyb}(S) = \sum_{A \in \mathcal{A}_S} \mathbb{P}(A) \cdot weight(A), \tag{4.2}$$

where $\mathbb{P}(A)$ is the probability of the arborescence A and $weight(A)$ is the sum of its branch lengths.

Example 9. Consider the phylogenetic network $\mathcal{N}_8$ depicted in Figure 21. We set $S = \{A, C\}$ and calculate the *hybrid phylogenetic net diversity* $PND^{hyb}_{\mathcal{N}_8}(S)$. There are three arborescences containing S and the root: A_1 with weight 5 and probability $\mathbb{P}(A_1) = \frac{1}{4}$, A_2 with weight 6 and probability $\mathbb{P}(A_2) = \frac{1}{2}$ and A_3 with weight 6 and probability $\mathbb{P}(A_3) = \frac{1}{4}$ (cf. Figure 21). Thus, we have

$$\begin{aligned} PND^{hyb}_{\mathcal{N}_8}(S) &= \frac{1}{4} \cdot 5 + \frac{1}{2} \cdot 6 + \frac{1}{4} \cdot 6 \\ &= \frac{23}{4}. \end{aligned}$$

Analogously, the *hybrid phylogenetic net diversity* can be calculated for all other subsets $S \subseteq X$ of taxa. Table 1 summarizes the results.

Remark. Note that the calculation of the *hybrid phylogenetic net diversity* involves the enumeration of all smallest arborescences spanning a subset $S \subseteq X$ of taxa and the root. In the worst case the number of such arborescences may be exponential in the number of reticulation nodes of the network. Consider, for example, the network $\mathcal{N}_1^*$ depicted in Figure 11 and fix the subset $S = \{A, E\} \subseteq X$ of taxa. Then there are $2^{r(\mathcal{N}_1^*)-1} + 1 = 17$ smallest arborescences spanning S and the root (i.e. arborescences

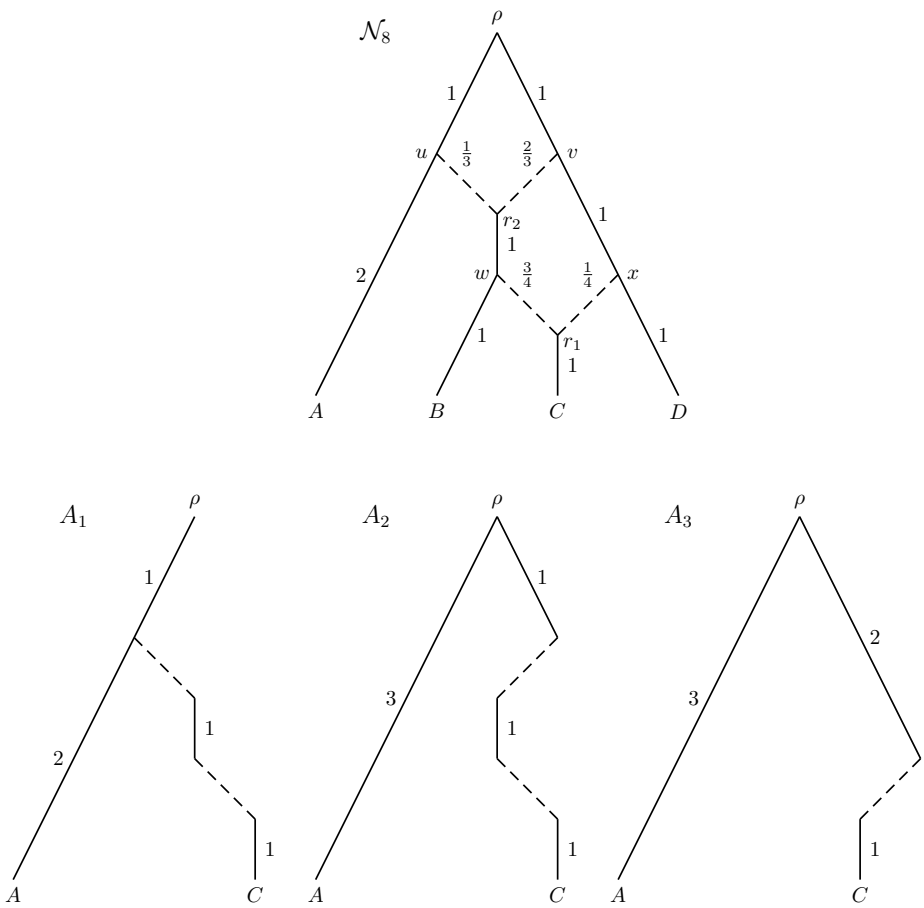


Fig. 21: Rooted phylogenetic network $\mathcal{N}_8$ on $X = \{A, B, C, D\}$ and smallest arborescences A_1, A_2 and A_3 containing $S = \{A, C\}$ and the root. A_1 has probability $\mathbb{P}(A_1) = \frac{1}{3} \cdot \frac{3}{4} \cdot 1 \cdot 1 \cdot 1 \cdot 1 = \frac{1}{4}$, A_2 has probability $\mathbb{P}(A_2) = \frac{2}{3} \cdot \frac{3}{4} \cdot 1 \cdot 1 \cdot 1 \cdot 1 = \frac{1}{2}$ and A_3 has probability $\mathbb{P}(A_3) = \frac{1}{4} \cdot 1 \cdot 1 \cdot 1 = \frac{1}{4}$.
Thus, A_2 is the *most likely arborescence* spanning S and the root.

that do not contain any additional edges than those required to connect S and the root).

A maximum likelihood approach

We now suggest a third approach, which considers the most likely arborescence containing the taxa in S and the root, and introduce the so-called *Maximum Likelihood phylogenetic net diversity.*

Definition 30 (Maximum Likelihood phylogenetic net diversity (PND^{ML}))**.** Let $\mathcal{N}$ be a rooted phylogenetic network on a taxon set X. Let $S \subseteq X$ be a subset of taxa and let $\mathcal{A}_{\mathcal{S}}$ be the set of all smallest arborescences spanning the taxa in S and the root, i.e. all arborescences (not necessarily of minimum cost) that span S and the root, but do not contain any additional nodes or edges than those necessary to connect the taxa in S and the root. Then we define the *Maximum Likelihood phylogenetic net diversity* $PND^{ML}(S)$ of S as the sum of branch lengths in the most likely arborescence A^* containing S and the root, i.e.

$$A^* = \operatorname*{argmax}_{A \in \mathcal{A}_{\mathcal{S}}} \mathbb{P}(A),$$

where $\mathbb{P}(A)$ is the probability of A. If the argmax is not unique, we choose the arborescence with minimum weight.[8]

Example 10. Consider the phylogenetic network $\mathcal{N}_8$ depicted in Figure 21. We set $S = \{A, C\}$ and calculate the *Maximum Likelihood phylogenetic net diversity* $PND^{ML}_{\mathcal{N}_8}(S)$. There are three arborescences containing S and the root: A_1 with weight 5 and probability $\mathbb{P}(A_1) = \frac{1}{4}$, A_2 with weight 6 and probability $\mathbb{P}(A_2) = \frac{1}{2}$ and A_3 with weight 6 and probability $\mathbb{P}(A_3) = \frac{1}{4}$ (cf. Figure 21). Thus, A_2 is the most likely arborescence and we have

$$PND^{ML}_{\mathcal{N}_8}(S) = weight(A_2) = 6.$$

Remark. Similar to the calculation of the *hybrid phylogenetic net diversity* the calculation of the *Maximum Likelihood phylogenetic net diversity* involves the enumeration of all smallest arborescences spanning a subset $S \subseteq X$ of taxa and the root, which, in the worst case, may be exponentially many (cf. Remark on page 44). It might be possible, though, to develop a greedy strategy for finding a smallest arborescence with maximum probability by choosing to include the reticulation edge with the highest probability

[8] Alternatively, we could arbitrarily choose one of the most likely arborescences. However, choosing the arborescence with minimum weight makes the results reproducible.

for each reticulation node that is encountered. However, we have not thought this idea through effectively, so we stick to the enumeration of all arborescences for the time being.

In a second approach, we now consider the set of phylogenetic X-trees displayed by a network $\mathcal{N}$ on X and define the *phylogenetic diversity* based on this set.

4.2. Embedded phylogenetic diversity

Phylogenetic networks are, mathematically, a generalization of phylogenetic trees, which can model reticulate evolution such as hybridization, i.e. evolution that is non-treelike.

Biologically, the genome of a hybrid species contains parts of the genome of both its ancestors. However, evolution at the nucleotide level rather than the genome level is still treelike, because a single nucleotide can always be traced back to one parent.

Therefore, we suggest to consider the set of embedded trees in a network (cf. Chapter 3.2) as an alternative approach to the generalization of *phylogenetic diversity* from trees to networks.

Definition 31 (Embedded phylogenetic diversity)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on a taxon set X and let $\mathsf{T}(\mathcal{N})$ be the (multi)set of all rooted phylogenetic X-trees displayed by $\mathcal{N}$. Then we use $PD^{*}_{\mathsf{T}(\mathcal{N})}(S)$ to denote the *embedded phylogenetic diversity* of a subset $S \subseteq X$ of taxa, where $*$ is one of the following functions $\min, \max, \sum, \varnothing$ and define

$$PD^{\min}_{\mathsf{T}(\mathcal{N})}(S) := \min_{\mathcal{T}\in\mathsf{T}(\mathcal{N})}\{PD_{\mathcal{T}}(S)\}, \tag{4.3}$$

$$PD^{\max}_{\mathsf{T}(\mathcal{N})}(S) := \max_{\mathcal{T}\in\mathsf{T}(\mathcal{N})}\{PD_{\mathcal{T}}(S)\}, \tag{4.4}$$

$$PD^{\sum}_{\mathsf{T}(\mathcal{N})}(S) := \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} PD_{\mathcal{T}}(S) \text{ and} \tag{4.5}$$

$$PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}(S) := \frac{1}{|\,\mathsf{T}(\mathcal{N})\,|}\sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} PD_{\mathcal{T}}(S), \tag{4.6}$$

where $|\,\mathsf{T}(\mathcal{N})\,|$ is the number of phylogenetic X-trees displayed by $\mathcal{N}$. If hybridization probabilities are given for $\mathcal{N}$, we also consider

$$PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(S) := \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} \mathbb{P}(\mathcal{T})\cdot PD_{\mathcal{T}}(S) \text{ and} \tag{4.7}$$

$$PD^{ML}_{\mathsf{T}(\mathcal{N})}(S) := PD_{\mathcal{T}^*}(S) \text{ with } \mathcal{T}^* = \operatorname*{argmax}_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} \mathbb{P}(\mathcal{T}), \tag{4.8}$$

where $\mathbb{P}(\mathcal{T})$ is the probability of $\mathcal{T}$ and $\mathcal{T}^*$ is the most likely embedded tree. If the argmax is not unique, we arbitrarily choose one of the embedded trees with maximum probability (cf. Definition 22).

Remarks.

- Note that $*$ can be replaced by other functions on the *phylogenetic diversity* of the trees in $\mathsf{T}(\mathcal{N})$, but we will only consider $\min, \max, \sum, \varnothing$ and $\varnothing_{hyb}$ as defined above.

- Furthermore, note that we will only consider phylogenetic X-trees as elements of $\mathsf{T}(\mathcal{N})$ and discard all other trees that may occur when decomposing the network into a set of trees (cf. Remark on page 20).

- As in the Remark on page 30 we assume $\alpha_{r_i} = \beta_{r_i} = \frac{1}{2}$ for all reticulation nodes r_i, $i = 1, \ldots, k$, if no explicit hybridization probabilities are given for a phylogenetic network with k reticulation nodes. In this case, all embedded trees have equal probability, namely $\mathbb{P}(\mathcal{T}) = \frac{1}{|\,\mathsf{T}(\mathcal{N})\,|}$ for all $\mathcal{T} \in \mathsf{T}(\mathcal{N})$, and the values of $PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}(S)$ and $PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(S)$ coincide for all subsets $S \subseteq X$ of taxa, because

$$\begin{aligned} PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(S) &= \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} \mathbb{P}(\mathcal{T}) \cdot PD_{\mathcal{T}}(S) \\ &= \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} \frac{1}{|\,\mathsf{T}(\mathcal{N})\,|} \cdot PD_{\mathcal{T}}(S) \\ &= \frac{1}{|\,\mathsf{T}(\mathcal{N})\,|} \cdot \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} PD_{\mathcal{T}}(S) \\ &= PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}(S). \end{aligned}$$

- Note that shifting from $PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}(S)$ to $PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(S)$ for phylogenetic networks with given hybridization probabilities means that the *weighted mean* instead of the *unweighted mean* of the *embedded phylogenetic diversity* of a subset of taxa $S \subseteq X$ is considered. However, there is no sufficient way to incorporate hybridization probabilities into the concepts of $PD^{\min}_{\mathsf{T}(\mathcal{N})}(S)$, $PD^{\max}_{\mathsf{T}(\mathcal{N})}(S)$ and $PD^{\sum}_{\mathsf{T}(\mathcal{N})}(S)$, because assigning probabilities to the trees in $\mathsf{T}(\mathcal{N})$ does not alter the minimum, maximum or sum of PD in those trees. For example,

$$PD^{\min}_{\mathsf{T}(\mathcal{N})}(S) = \min_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} PD_{\mathcal{T}}(S)$$

  is uniquely determined, regardless of whether the corresponding tree (not necessarily unique) that yields this value, has a high or low probability. The same holds

for the maximum. For $PD^{\sum}_{\mathsf{T}(\mathcal{N})}(S)$ the values of $PD(S)$ for each tree are added, again, regardless of the trees' probabilities.

Example 11. Consider the rooted binary phylogenetic network $\mathcal{N}_8$ and its embedded trees $\mathcal{T}_1', \ldots, \mathcal{T}_4'$ depicted in Figure 14. Note that we have $\mathbb{P}(\mathcal{T}_1') = \frac{1}{6}$, $\mathbb{P}(\mathcal{T}_2') = \frac{1}{2}$, $\mathbb{P}(\mathcal{T}_3') = \frac{1}{4}$ and $\mathbb{P}(\mathcal{T}_4') = \frac{1}{12}$. Thus, $\mathcal{T}_2'$ is the most likely embedded tree.
Exemplarily, we set $S = \{A, B\}$ and calculate all versions of the *embedded phylogenetic diversity* for S.

$$\begin{aligned}
PD^{\min}_{\mathsf{T}(\mathcal{N}_8)}(S) &= \min_{\mathcal{T}\in\mathsf{T}(\mathcal{N}_8)} PD_{\mathcal{T}}(S) \\
&= \min\{PD_{\mathcal{T}_1'}(S), PD_{\mathcal{T}_2'}(S), PD_{\mathcal{T}_3'}(S), PD_{\mathcal{T}_4'}(S)\} \\
&= \min\{6, 6, 5, 5\} \\
&= 5,
\end{aligned}$$

$$\begin{aligned}
PD^{\max}_{\mathsf{T}(\mathcal{N}_8)}(S) &= \max_{\mathcal{T}\in\mathsf{T}(\mathcal{N}_8)} PD_{\mathcal{T}}(S) \\
&= \max\{PD_{\mathcal{T}_1'}(S), PD_{\mathcal{T}_2'}(S), PD_{\mathcal{T}_3'}(S), PD_{\mathcal{T}_4'}(S)\} \\
&= \max\{6, 6, 5, 5\} \\
&= 6,
\end{aligned}$$

$$\begin{aligned}
PD^{\sum}_{\mathsf{T}(\mathcal{N}_8)}(S) &= \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N}_8)} PD_{\mathcal{T}}(S) \\
&= \sum\{PD_{\mathcal{T}_1'}(S), PD_{\mathcal{T}_2'}(S), PD_{\mathcal{T}_3'}(S), PD_{\mathcal{T}_4'}(S)\} \\
&= 6 + 6 + 5 + 5 \\
&= 22,
\end{aligned}$$

$$\begin{aligned}
PD^{\varnothing}_{\mathsf{T}(\mathcal{N}_8)}(S) &= \frac{1}{|\mathsf{T}(\mathcal{N}_8)|} \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N}_8)} PD_{\mathcal{T}}(S) \\
&= \frac{1}{4}\Big(PD_{\mathcal{T}_1'}(S) + PD_{\mathcal{T}_2'}(S) + PD_{\mathcal{T}_3'}(S) + PD_{\mathcal{T}_4'}(A)\Big) \\
&= \frac{1}{4}\Big(6 + 6 + 5 + 5\Big) \\
&= \frac{11}{2},
\end{aligned}$$

$$
\begin{aligned}
PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_8)}(S) &= \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N}_8)} \mathbb{P}(\mathcal{T})\cdot PD_{\mathcal{T}}(S) \\
&= \mathbb{P}(\mathcal{T}_1')\cdot PD_{\mathcal{T}_1'}(S) + \mathbb{P}(\mathcal{T}_2')\cdot PD_{\mathcal{T}_2'}(S) + \mathbb{P}(\mathcal{T}_3')\cdot PD_{\mathcal{T}_3'}(S) + \mathbb{P}(\mathcal{T}_4')\cdot PD_{\mathcal{T}_4'}(S) \\
&= \frac{1}{6}\cdot 6 + \frac{1}{2}\cdot 6 + \frac{1}{4}\cdot 5 + \frac{1}{12}\cdot 5 \\
&= \frac{17}{3},
\end{aligned}
$$

and

$$
\begin{aligned}
PD^{ML}_{\mathsf{T}(\mathcal{N}_8)}(S) &= PD_{\mathcal{T}_2'}(S) \\
&= 6.
\end{aligned}
$$

Analogously, the *embedded phylogenetic diversity* can be calculated for all other subsets $S \subseteq X$ of taxa. Table 1 summarizes the results.

Table 1: *Phylogenetic net diversity*, *embedded phylogenetic diversity* and *LSA associated phylogenetic diversity* for the rooted phylogenetic network $\mathcal{N}_8$

S	PND	PND^{hyb}	PND^{ML}	$PD^{\min}_{\mathsf{T}(\mathcal{N}_8)}$	$PD^{\max}_{\mathsf{T}(\mathcal{N}_8)}$	$PD^{\sum}_{\mathsf{T}(\mathcal{N}_8)}$	$PD^{\varnothing}_{\mathsf{T}(\mathcal{N}_8)}$	$PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_8)}$	$PD^{ML}_{\mathsf{T}(\mathcal{N}_8)}$	PD^{LSA}	$PD^{LSA_{hyb}}$	$PD^{LSA_{ML}}$
$\emptyset$	0	0	0	0	0	0	0	0	0	0	0	0
$\{A\}$	3	3	3	3	3	12	3	3	3	3	3	3
$\{B\}$	3	3	3	3	3	12	3	3	3	3	3	3
$\{C\}$	3	3	3	3	3	12	3	3	3	3	3	3
$\{D\}$	3	3	3	3	3	12	3	3	3	3	3	3
$\{A,B\}$	5	$\frac{17}{3}$	6	5	6	22	$\frac{11}{2}$	$\frac{17}{3}$	6	6	6	6
$\{A,C\}$	5	$\frac{23}{4}$	6	5	6	23	$\frac{23}{4}$	$\frac{23}{4}$	6	6	6	6
$\{A,D\}$	6	6	6	6	6	24	6	6	6	6	6	6
$\{B,C\}$	4	$\frac{13}{3}$	4	4	6	19	$\frac{19}{4}$	$\frac{13}{3}$	4	6	6	6
$\{B,D\}$	5	$\frac{16}{3}$	5	5	6	22	$\frac{11}{2}$	$\frac{16}{3}$	5	6	6	6
$\{C,D\}$	4	5	5	4	6	19	$\frac{19}{4}$	5	5	6	6	6
$\{A,B,C\}$	6	7	7	6	8	29	$\frac{29}{4}$	7	7	9	9	9
$\{A,B,D\}$	8	8	8	8	8	32	8	8	8	9	9	9
$\{A,C,D\}$	7	$\frac{31}{4}$	8	7	8	30	$\frac{15}{2}$	$\frac{31}{4}$	8	9	9	9
$\{B,C,D\}$	6	$\frac{19}{3}$	6	6	7	26	$\frac{13}{2}$	$\frac{19}{3}$	6	9	9	9
$\{A,B,C,D\}$	9	9	9	9	9	36	9	9	9	12	12	12

Relationship between the phylogenetic net diversity and the embedded phylogenetic diversity

Comparing the *phylogenetic net diversity* PND and the *minimum embedded phylogenetic diversity* $PD^{\min}_{\mathsf{T}(\mathcal{N})}$ for a subset $S \subseteq X$ of taxa, we see that they use a similar principle. While $PND(S)$ is defined as the weight of a minimum-cost arborescence spanning S and the root in a network $\mathcal{N}$, $PD^{\min}_{\mathsf{T}(\mathcal{N})}(S)$ is defined as the weight of a minimum-cost arborescence spanning S and the root in the (multi)set $\mathsf{T}(\mathcal{N})$ of phylogenetic X-trees displayed by $\mathcal{N}$. Thus, the two measures are related, but in general they are not identical.

Consider, for example, the rooted phylogenetic network $\mathcal{N}_2$ depicted in Figure 6 and set $S = \{A, B, C, D\}$. Then, we have

$$PD^{\min}_{\mathsf{T}(\mathcal{N}_2)}(S) = 9,$$

while

$$PND_{\mathcal{N}_2}(S) = 8.$$

However, we have the following relationship between PND and $PD^{\min}_{\mathsf{T}(\mathcal{N})}$:

Proposition 2. *Let $\mathcal{N}$ be a binary rooted phylogenetic network on a taxon set X with k reticulation nodes and let $\mathsf{T}(\mathcal{N})$ be the set of phylogenetic X-trees displayed by $\mathcal{N}$.*

1. *It is*
$$PND(S) \leq PD^{\min}_{\mathsf{T}(\mathcal{N})}(S) \tag{4.9}$$
for all subsets $S \subseteq X$ of taxa.

2. *If $|\,\mathsf{T}(\mathcal{N})\,| = 2^k$, i.e. if all combinations of removing one reticulation edge for each reticulation node and suppressing nodes of both indegree 1 and outdegree 1 result in a phylogenetic X-tree, we have*
$$PND(S) = PD^{\min}_{\mathsf{T}(\mathcal{N})}(S). \tag{4.10}$$

Proof. Let $\mathcal{N}$ be a binary rooted phylogenetic network with root ρ, taxon set X and k reticulation nodes. Let $\mathsf{T}(\mathcal{N})$ be the set of embedded trees and let $R(\mathcal{N}) = \{r \,|\, r \text{ is a reticulation node of } \mathcal{N}\}$ be the set of reticulation nodes of N.

1. We show $PD^{\min}_{\mathsf{T}(\mathcal{N})}(S) \geq PND(S)$.
For every $\mathcal{T} \in \mathsf{T}(\mathcal{N})$ the *phylogenetic diversity* of a subset $S \subseteq X$ of taxa is defined as the sum of branch lengths in the smallest arborescence spanning the taxa in S and the root. Clearly, the weight of any such arborescence cannot be

smaller than the weight of a minimum cost arborescence spanning S and the root in $\mathcal{N}$ (all $\mathcal{T} \in \mathsf{T}(\mathcal{N})$ are 'subgraphs' of $\mathcal{N}$, thus, any smallest arborescence spanning S and the root in a displayed tree $\mathcal{T} \in \mathsf{T}(\mathcal{N})$, can also be found in $\mathcal{N}$).[9] In particular, we have

$$\min_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} \{PD_{\mathcal{T}}(S)\} = PD^{\min}_{\mathsf{T}(\mathcal{N})}(S) \geq PND(S).$$

2. Now, suppose that $|\,\mathsf{T}(\mathcal{N})\,| = 2^k$. We want to show that $PND(S) = PD^{\min}_{\mathsf{T}(\mathcal{N})}(S)$. As we have $PND(S) \leq PD^{\min}_{\mathsf{T}(\mathcal{N})}(S)$ (Equation (4.9)), it suffices to show $PND(S) \geq PD^{\min}_{\mathsf{T}(\mathcal{N})}(S)$.
   Let A_S be the minimum cost arborescence spanning S and the root in $\mathcal{N}$. By definition of an arborescence there is exactly one directed path from the root ρ to any other vertex $v \in V(A_S)$. This implies that A_S contains at most one reticulation edge for each reticulation node $r \in R(\mathcal{N})$, but never both reticulation edges directed into $r \in R(\mathcal{N})$. If we now suppress nodes of both indegree 1 and outdegree 1 in A_S and add the weights of the edges which are merged into one edge by doing so, we retrieve a directed acyclic graph A'_S, which contains the taxa in S and whose weight equals the weight of A_S. By the construction of A'_S, however, A'_S must be a sub-arborescence of some embedded tree $\mathcal{T}_{A_S} \in \mathsf{T}(\mathcal{N})$, where the set of embedded trees is obtained by deleting one of the reticulation edges for each reticulation node and suppressing the resulting nodes of indegree 1 and outdegree 1, and every combination of doing so results in a phylogenetic X-tree (because we have assumed $|\,\mathsf{T}(\mathcal{N})\,| = 2^k$). Thus, by definition of PD for trees, the weight of A_S equals $PD_{\mathcal{T}_{A_S}}(S)$ and as $\mathcal{T}_{A_S}$ is embedded in $\mathcal{N}$ we have

$$PND(S) = PD_{\mathcal{T}_{A_S}}(S) \geq \min_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} \{PD_{\mathcal{T}}(S)\} = PD^{\min}_{\mathsf{T}(\mathcal{N})}(S).$$

   Combining the above, we have $PND(S) = PD^{\min}_{\mathsf{T}(\mathcal{N})}(S)$ as claimed.

□

Remarks.

- Part 2 of Proposition 2 does not hold for general networks, because the arborescence A'_S (see Proof of part 2) does not necessarily have to be a sub-arborescence of an embedded tree, which is crucial for the proof of Equation (4.10). Consider,

[9] Formally, we have to re-establish the nodes of in- and outdegree 1 that were removed during the construction of $\mathcal{T} \in \mathsf{T}(\mathcal{N})$ to make $\mathcal{T}$ a subgraph of $\mathcal{N}$. However, this does not affect the weight of any arborescence spanning S and the root in $\mathcal{T}$, but it becomes obvious that any such arborescence can then be found in $\mathcal{N}$ as well.

for example, the rooted phylogenetic network $\mathcal{N}_2$ depicted in Figure 6. We have already seen that $PND_{\mathcal{N}_2}(\{A, B, C, D\}) = 8$. In this case the minimum cost arborescence spanning all taxa and the root is a 'subgraph' of $\mathcal{T}_4'$ (cf. Figure 6), but $\mathcal{T}_4'$ is not a phylogenetic X-tree and thus, not displayed by $\mathcal{N}_2$.
However, if all combinations of removing one reticulation edge for each reticulation node and suppressing nodes of both indegree 1 and outdegree 1 result in a phylogenetic X-tree, the arborescence A_S' has to be a sub-arborescence of some embedded tree, and thus, $PND(S)$ and $PD^{\min}_{\mathsf{T}(\mathcal{N})}(S)$ coincide for all subsets $S \subseteq X$ of taxa.

- Note, however, that the *phylogenetic net diversity* $PND(S)$ of a subset $S \subseteq X$ of taxa can still be calculated by decomposing the network into phylogenetic trees in the case $|\,\mathsf{T}(\mathcal{N})\,| \leq 2^k$, if we do not only consider the set of embedded phylogenetic X-trees, but also the set of phylogenetic X'-trees with $X \neq X'$, that may occur (e.g. phylogenetic trees, where internal nodes of $\mathcal{N}$ have become leaves). Let $\overline{\mathsf{T}(\mathcal{N})}$ be this extended (multi)set of phylogenetic trees obtained from $\mathcal{N}$. Then we have
$$PND(S) = PD^{\min}_{\overline{\mathsf{T}(\mathcal{N})}}(S)$$
for all $S \subseteq X$, because any minimum cost arborescence A_S' will be a sub-arborescence of some tree $\mathcal{T}' \in \overline{\mathsf{T}(\mathcal{N})}$ and we can apply part 2 of the above proof and thus, of Proposition 2.

Relationship between the hybrid phylogenetic diversity and the embedded phylogenetic diversity

If we now compare the *hybrid phylogenetic net diversity* $PND^{hyb}(S)$ and the *hybrid average embedded phylogenetic diversity* $PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(S)$ of a subset $S \subseteq X$ of taxa, we see that these values, again, follow a related principle. While $PND^{hyb}(S)$ considers all smallest arborescences spanning S and the root in $\mathcal{N}$, $PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(S)$ considers the smallest arborescence spanning S and the root in each of the phylogenetic trees displayed by $\mathcal{N}$. Thus, again, the two values are related, but do not coincide in general.

Consider, for example, the rooted phylogenetic network $\mathcal{N}_2'$ depicted in Figure 16 and set $S = \{A, B, C, D\}$. There are three embedded trees in $\mathsf{T}(\mathcal{N}_2')$, but four smallest arborescences spanning S and the root in $\mathcal{N}_2'$ (the first three arborescences coincide with $\mathcal{T}_1', \mathcal{T}_2'$ and $\mathcal{T}_3'$, while the fourth arborescence is a subgraph of $\mathcal{T}_4'$ (after suppressing nodes of in- and outdegree 1 in the arborescences)). Thus,

$$PD^{\varnothing hyb}_{\mathsf{T}(\mathcal{N}_2')}(S) = \frac{1}{6} \cdot 9 + \frac{1}{3} \cdot 9 + \frac{1}{2} \cdot 9 = 9,$$

while

$$PND^{hyb}_{\mathcal{N}_2'}(S) = \frac{1}{12} \cdot 9 + \frac{1}{6} \cdot 9 + \frac{1}{4} \cdot 9 + \frac{1}{2} \cdot 8 = \frac{17}{2}.$$

Considering not only the set of embedded phylogenetic X-trees $\mathsf{T}(\mathcal{N}_2')$, but the extended set of embedded trees $\overline{\mathsf{T}(\mathcal{N}_2')} = \{\mathcal{T}_1', \mathcal{T}_2', \mathcal{T}_3', \mathcal{T}_4'\}$ (cf. Remark on page 54), we have

$$\begin{aligned} PD^{hyb}_{\overline{\mathsf{T}(\mathcal{N}_2')}}(S) &= \frac{1}{12} \cdot 9 + \frac{1}{6} \cdot 9 + \frac{1}{4} \cdot 9 + \frac{1}{2} \cdot 8 = \frac{17}{2} \\ &= PND^{hyb}_{\mathcal{N}_2'}(S), \end{aligned}$$

thus, in this case the values coincide.

Proposition 3. *Let $\mathcal{N}$ be a binary rooted phylogenetic network on a taxon set X with k reticulation nodes and let $\mathsf{T}(\mathcal{N})$ be the set of phylogenetic X-trees displayed by $\mathcal{N}$.*

1. *If $|\,\mathsf{T}(\mathcal{N})\,| \leq 2^k$, we have*

$$PND^{hyb}(S) = PD^{\varnothing hyb}_{\overline{\mathsf{T}(\mathcal{N})}}(S), \tag{4.11}$$

   where $\overline{\mathsf{T}(\mathcal{N})}$ is the (multi)set of all trees (not necessarily phylogenetic X-trees) that can be obtained from $\mathcal{N}$ by removing one of the reticulation edges for each reticulation node and suppressing the resulting nodes of indegree 1 and outdegree 1.

2. *If $|\,\mathsf{T}(\mathcal{N})\,| = 2^k$, i.e. if all combinations of removing one reticulation edge for each reticulation node and suppressing nodes of both indegree 1 and outdegree 1 result in a phylogenetic X-tree, we have:*

$$PND^{hyb}(S) = PD^{\varnothing hyb}_{\mathsf{T}(\mathcal{N})}(S). \tag{4.12}$$

Remark. If $|\,\mathsf{T}(\mathcal{N})\,| = 2^k$ in case 2 of Proposition 3, obviously, $\overline{\mathsf{T}(\mathcal{N})} = \mathsf{T}(\mathcal{N})$. Thus, case 2 is a special case of case 1.

Idea of Proof. The reasoning in this proof is similar to the proof of Proposition 2, so we only give the idea.
Let $\mathcal{N}$ be a binary rooted phylogenetic network on X and let $S \subseteq X$ be a subset of taxa.

1. The *hybrid phylogenetic net diversity* of a subset S of taxa is defined as the weighted average of weights of smallest arborescences spanning S and the root,

i.e.

$$PND^{hyb}(S) = \sum_{A \in \mathcal{A}_S} \mathbb{P}^{\mathcal{N}}(A) \cdot weight(A),$$

where $\mathbb{P}^{\mathcal{N}}(A)$ is the probability of A in $\mathcal{N}$ and $\mathcal{A}_S = \{A_1, \ldots, A_l\}$ is the set of smallest arborescences spanning S and the root in $\mathcal{N}$.

The *average embedded hybrid phylogenetic diversity* of S, on the other, hand is defined as

$$PD^{\varnothing hyb}_{\overline{\mathsf{T}(\mathcal{N})}}(S) = \sum_{\mathcal{T} \in \overline{\mathsf{T}(\mathcal{N})}} \mathbb{P}(\mathcal{T}) \cdot PD_{\mathcal{T}}(S).$$

Now we can break down this sum in the following way:

$$\begin{aligned}
PD^{\varnothing hyb}_{\overline{\mathsf{T}(\mathcal{N})}}(S) &= \sum_{\mathcal{T} \in \overline{\mathsf{T}(\mathcal{N})}} \mathbb{P}(\mathcal{T}) \cdot PD_{\mathcal{T}}(S) \\
&= \sum_{\substack{\mathcal{T} \in \overline{\mathsf{T}(\mathcal{N})}: \\ \mathcal{T} \text{ contains } A_1}} \mathbb{P}(\mathcal{T}) \cdot PD_{\mathcal{T}}(S) + \ldots + \sum_{\substack{\mathcal{T} \in \overline{\mathsf{T}(\mathcal{N})}: \\ \mathcal{T} \text{ contains } A_l}} \mathbb{P}(\mathcal{T}) \cdot PD_{\mathcal{T}}(S) \\
&= \sum_{\substack{\mathcal{T} \in \overline{\mathsf{T}(\mathcal{N})}: \\ \mathcal{T} \text{ contains } A_1}} \mathbb{P}(\mathcal{T}) \cdot weight(A_1) + \ldots + \sum_{\substack{\mathcal{T} \in \overline{\mathsf{T}(\mathcal{N})}: \\ \mathcal{T} \text{ contains } A_l}} \mathbb{P}(\mathcal{T}) \cdot weight(A_l) \\
&= \sum_{A \in \mathcal{A}_S} \mathbb{P}^{\overline{\mathsf{T}(\mathcal{N})}}(A) \cdot weight(A),
\end{aligned}$$

with

$$\mathbb{P}^{\overline{\mathsf{T}(\mathcal{N})}}(A) := \sum_{\substack{\mathcal{T} \in \overline{\mathsf{T}(\mathcal{N})}: \\ \mathcal{T} \text{ contains } A}} \mathbb{P}(\mathcal{T}),$$

i.e. $\mathbb{P}^{\overline{\mathsf{T}(\mathcal{N})}}(A)$ denotes the probability of the arborescence A in the extended set of embedded trees.[10]

It remains to show that we have $\mathbb{P}^{\overline{\mathsf{T}(\mathcal{N})}}(A) = \mathbb{P}^{\mathcal{N}}(A)$ for all arborescences $A \in \mathcal{A}_S$. The probability $\mathbb{P}^{\mathcal{N}}(A)$ of an arborescence in A in $\mathcal{N}$ is calculated as the product of the probabilities of its reticulation edges.

When calculating $\mathbb{P}^{\overline{\mathsf{T}(\mathcal{N})}}(A)$ we have to consider all trees in $\overline{\mathsf{T}(\mathcal{N})}$ and sum up their probabilities. If a tree $\mathcal{T}$ in $\overline{\mathsf{T}(\mathcal{N})}$ contains the arborescence A, it has to contain the reticulation edges present in A, but it might contain additional reticulation

[10] Formally, we have to re-establish the nodes of in- and outdegree 1 that were removed during the construction of $\mathcal{T} \in \overline{\mathsf{T}(\mathcal{N})}$ for all $\mathcal{T} \in \overline{\mathsf{T}(\mathcal{N})}$. Then we can say that a tree $\mathcal{T}$ 'contains' an arborescence $A \in \mathcal{A}_S$ or a certain reticulation edge. Thus, for the rest of the proof we assume all embedded trees $\mathcal{T} \in \overline{\mathsf{T}(\mathcal{N})}$ to have regained the nodes of in- and outdegree 1 that were removed during their construction. Note however, that this does not influence the *weight* of any arborescence in a tree $\mathcal{T} \in \overline{\mathsf{T}(\mathcal{N})}$.

edges. However,

$$\begin{aligned}
\mathbb{P}^{\overline{\mathsf{T}(\mathcal{N})}}(A) &= \sum_{\substack{\mathcal{T}\in\overline{\mathsf{T}(\mathcal{N})}:\\ \mathcal{T}\text{ contains }A}} \mathbb{P}(\mathcal{T}) \\
&= \sum_{\substack{\mathcal{T}\in\overline{\mathsf{T}(\mathcal{N})}:\\ \mathcal{T}\text{ contains }A}} \prod_{e\in\mathcal{T}} \mathbb{P}(e) \\
&= \sum_{\substack{\mathcal{T}\in\overline{\mathsf{T}(\mathcal{N})}:\\ \mathcal{T}\text{ contains }A}} \underbrace{\prod_{e\in A} \mathbb{P}(e)}_{\substack{\text{reticulation}\\ \text{edges in }A}} \underbrace{\prod_{e\in\mathcal{T}\setminus A} \mathbb{P}(e)}_{\substack{\text{additional}\\ \text{reticulation edges}}} \\
&= \underbrace{\prod_{e\in A} \mathbb{P}(e)}_{=\mathbb{P}^{\mathcal{N}}(A)} \underbrace{\sum_{\substack{\mathcal{T}\in\overline{\mathsf{T}(\mathcal{N})}:\\ \mathcal{T}\text{ contains }A}} \prod_{e\in\mathcal{T}\setminus A} \mathbb{P}(e)}_{=1} \\
&= \mathbb{P}^{\mathcal{N}}(A).
\end{aligned}$$

Thereby

$$\sum_{\substack{\mathcal{T}\in\overline{\mathsf{T}(\mathcal{N})}:\\ \mathcal{T}\text{ contains }A}} \prod_{e\in\mathcal{T}\setminus A} \mathbb{P}(e) = 1,$$

because the sum runs over all trees $\mathcal{T}$ in $\overline{\mathsf{T}(\mathcal{N})}$ that contain A and considers the portion of their probability that does not arise from the reticulation edges in A, but from additional reticulation edges. However, as $\overline{\mathsf{T}(\mathcal{N})}$ contains all trees constructed by all possible combinations of removing one reticulation edge for each reticulation node and deleting nodes of indegree 1 and outdegree 1, the sum over these portions must equal 1, because the additional reticulation edges complement each other (cf. Figure 22).
Summarizing the above, we have

$$PND^{hyb}(S) = PD^{\varnothing hyb}_{\overline{\mathsf{T}(\mathcal{N})}}(S)$$

as claimed.

2. Special case of 1. with $\overline{\mathsf{T}(\mathcal{N})} = \mathsf{T}(\mathcal{N})$.

□

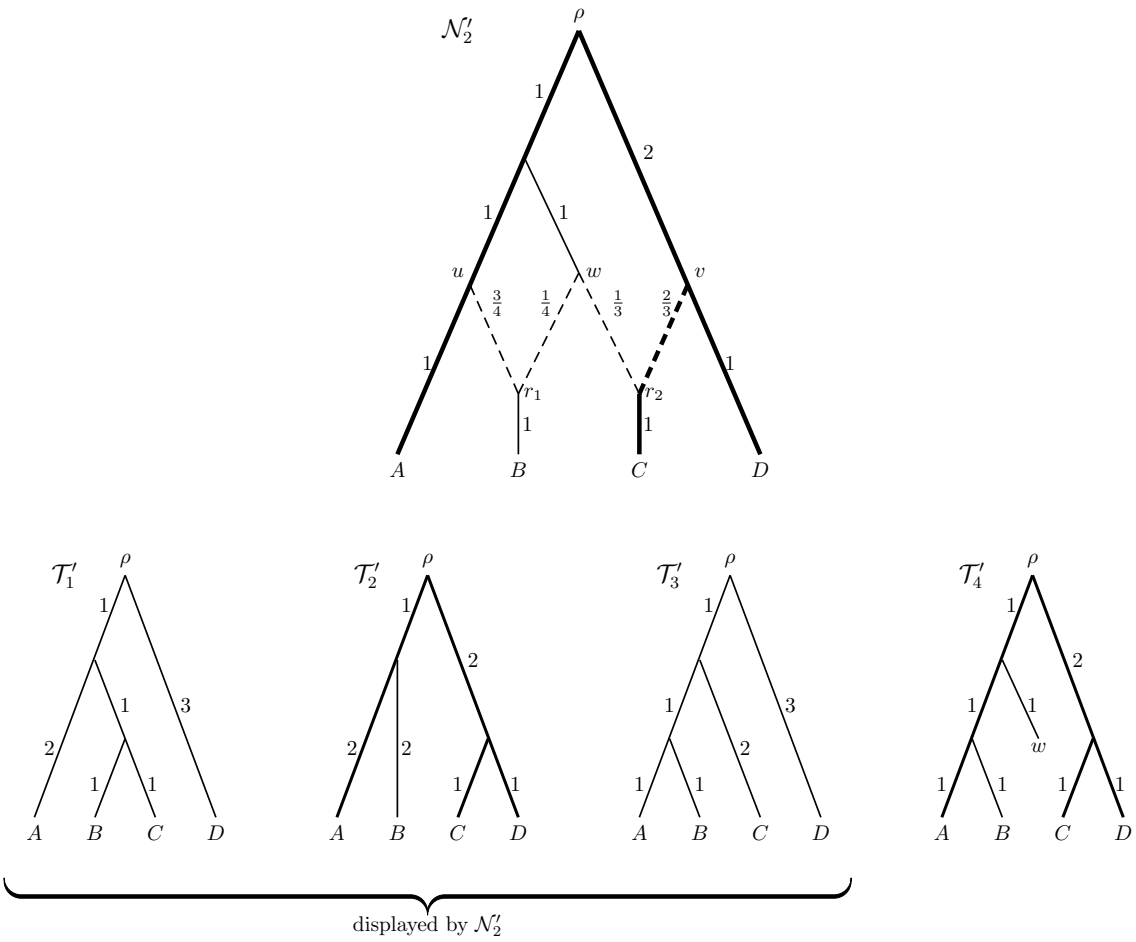


Fig. 22: The probability of the arborescence $\mathcal{A}$ spanning $S = \{A, C, D\}$ and the root in $\mathcal{N}_2'$ (bold lines) has probability $\mathbb{P}^{\mathcal{N}_2'}(\mathcal{A}) = \frac{2}{3}$.
When considering $\overline{\mathsf{T}(\mathcal{N}_2')}$, we see that $\mathcal{A}$ is contained in $\mathcal{T}_2'$ and $\mathcal{T}_4'$. We have $\mathbb{P}(\mathcal{T}_2') = \frac{1}{4} \cdot \frac{2}{3} = \frac{1}{6}$ and $\mathbb{P}(\mathcal{T}_4') = \frac{3}{4} \cdot \frac{2}{3} = \frac{1}{2}$. Thus, $\mathbb{P}^{\overline{\mathsf{T}(\mathcal{N}_2')}}(\mathcal{A}) = \frac{1}{6} + \frac{1}{2} = \frac{2}{3} = \mathbb{P}^{\mathcal{N}_2'}(\mathcal{A})$. Note that $\mathcal{T}_2'$ and $\mathcal{T}_4'$ contain the former reticulation edge (v, r_2) present in $\mathcal{N}_2'$, but both $\mathcal{T}_2'$ and $\mathcal{T}_4'$ contain an additional former reticulation edge, respectively (in $\mathcal{T}_2'$ the edge (w, r_1) was kept and in $\mathcal{T}_4'$ the edge (u, r_1)). The probabilities of these edges, however, complement each other, because they are directed into the same reticulation node.

Remark. If $|\,\mathsf{T}(\mathcal{N})\,| < 2^k$ for a phylogenetic network $\mathcal{N}$ on $X = \{A, B, C, D\}$ with k reticulation nodes and we are not considering the extended (multi)set $\overline{\mathsf{T}(\mathcal{N})}$ of embedded trees, but the set of $\mathsf{T}(\mathcal{N})$ of phylogenetic X-trees displayed by $\mathcal{N}$, we cannot predict the relationship of $PND^{hyb}(S)$ and $PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(S)$. We might either have $PND^{hyb}(S) \leq PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(S), PND^{hyb}(S) = PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(S)$ or $PND^{hyb}(S) \geq PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(S)$. Consider, for example, the rooted phylogenetic network $\mathcal{N}_2'$ with two reticulation nodes depicted in Figure 16.

- For $S = \{A, B\}$ we have

$$\begin{aligned} PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_2')}(S) &= \frac{1}{6} \cdot 5 + \frac{1}{3} \cdot 5 + \frac{1}{2} \cdot 4 \\ &= \frac{9}{2}, \end{aligned}$$

  but

$$\begin{aligned} PND^{hyb}(S) &= \frac{3}{4} \cdot 4 + \frac{1}{4} \cdot 5 \\ &= \frac{17}{4}, \end{aligned}$$

  thus, $PND^{hyb}(S) \leq PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_2')}(S)$.

- For $S = \{A, C\}$ we have

$$\begin{aligned} PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_2')}(S) &= \frac{1}{6} \cdot 5 + \frac{1}{3} \cdot 6 + \frac{1}{2} \cdot 5 \\ &= \frac{16}{3}, \end{aligned}$$

  but

$$\begin{aligned} PND^{hyb}(S) &= \frac{1}{3} \cdot 5 + \frac{2}{3} \cdot 6 \\ &= \frac{17}{3}, \end{aligned}$$

  thus, $PND^{hyb}(S) \geq PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_2')}(S)$.

- For $S = \{A, D\}$ we have

$$\begin{aligned} PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_2')}(S) &= \frac{1}{6} \cdot 6 + \frac{1}{3} \cdot 6 + \frac{1}{2} \cdot 6 \\ &= 6, \end{aligned}$$

and

$$PND^{hyb}(S) = 1 \cdot 6 = 6,$$

thus, $PND^{hyb}(S) = PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}'_2)}(S)$.

Remark. Summarizing the above, we see that the *phylogenetic net diversity* PND and the *embedded phylogenetic diversity* $PD^*_{\mathsf{T}(\mathcal{N})}$ to some extent follow similar ideas. Recall, however, that the calculation of the *phylogenetic net diversity* PND was an instance of the directed Steiner tree problem and thus, an NP-hard problem. If we recapitulate on how $PD^*_{\mathsf{T}(\mathcal{N})}(S)$ is calculated for a subset $S \subseteq X$ of taxa, we see that several steps are involved:

1. Determination of the set of displayed trees $\mathsf{T}(\mathcal{N})$.
2. Calculation of $PD_{\mathcal{T}}(S)$ for all $\mathcal{T} \in \mathsf{T}(\mathcal{N})$.
3. Calculation of $PD^*_{\mathsf{T}(\mathcal{N})}(S)$.

Steps 2 and 3 can be achieved in linear time, step 1, however, makes the approach NP-hard as well (cf. Linz et al. [25]).

Thus, the calculation may be infeasible, especially for a growing number of reticulation nodes in $\mathcal{N}$.

Therefore, we now suggest a third way of generalizing the concept of *phylogenetic diversity* from trees to networks, which makes use of the *LSA tree* introduced in Chapter 3.3.

4.3. LSA associated phylogenetic diversity

The lowest stable ancestor tree introduced in Chapter 3.3 can be seen as a way to summarize the treelike content of a phylogenetic network, on which all its embedded trees agree, without explicitly having to consider these trees. We therefore suggest to use the *LSA tree* associated with a phylogenetic network as a third approach towards the generalization of *phylogenetic diversity* from trees to networks.

Definition 32 (LSA associated phylogenetic diversity)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on some taxon set X. Let $S \subseteq X$ be a subset of taxa. Then we define

$$PD^{LSA}(S) := PD_{\mathcal{T}_{LSA}(\mathcal{N})}(S), \tag{4.13}$$

where $PD_{\mathcal{T}_{LSA}(\mathcal{N})}(S)$ is the *phylogenetic diversity* of S in the *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N})$ associated with $\mathcal{N}$. If hybridization probabilities are given for $\mathcal{N}$, we also consider

$$PD^{LSA_{hyb}}(S) \coloneqq PD_{\mathcal{T}^{hyb}_{LSA}(\mathcal{N})}(S) \text{ and} \tag{4.14}$$

$$PD^{LSA_{ML}}(S) \coloneqq PD_{\mathcal{T}^{ML}_{LSA}(\mathcal{N})}(S), \tag{4.15}$$

where $PD_{\mathcal{T}^{hyb}_{LSA}(\mathcal{N})}(S)$ is the *phylogenetic diversity* of S in the *hybrid LSA tree* $\mathcal{T}^{hyb}_{LSA}(\mathcal{N})$ associated with $\mathcal{N}$ and $PD_{\mathcal{T}^{ML}_{LSA}(\mathcal{N})}(S)$ is the *phylogenetic diversity* of S in the *Maximum Likelihood LSA tree* $\mathcal{T}^{ML}_{LSA}(\mathcal{N})$ associated with $\mathcal{N}$.
We will call all three versions *LSA associated phylogenetic diversity* and use the superscript to indicate which version of the *LSA tree* was used.

Remark. Note that the *LSA associated phylogenetic diversity* $PD^{LSA}_{\mathcal{N}}(S)$ and the *hybrid LSA associated phylogenetic diversity* $PD^{LSA_{hyb}}_{\mathcal{N}}(S)$ coincide for all subsets $S \subseteq X$ of taxa, if no hybridization probabilities are given. In this case, we assume all paths $P \in \mathcal{P}_r$ between a reticulation node r and its lowest stable ancestor $lsa(r)$ to be equally likely, thus, $\mathbb{P}(P) = \frac{1}{|\mathcal{P}_r|}$, where $\mathcal{P}_r$ is the set of all $lsa(r) - r-$paths in $\mathcal{N}$. The length of the edge $e = (lsa(r), r)$ is then calculated as

$$\begin{aligned} length(e = (lsa(r), r)) &= \sum_{P \in \mathcal{P}_r} \mathbb{P}(P) \cdot length(P) \\ &= \sum_{P \in \mathcal{P}_r} \frac{1}{|\mathcal{P}_r|} \cdot length(P) \\ &= \frac{1}{|\mathcal{P}_r|} \sum_{P \in \mathcal{P}_r} length(P), \end{aligned}$$

and thus, the *hybrid LSA tree* $\mathcal{T}^{hyb}_{LSA}(\mathcal{N})$ and the *LSA tree* $\mathcal{T}_{LSA}$ coincide (cf. Equation (3.1)). Subsequently, the *LSA associated phylogenetic diversity* $PD^{LSA}_{\mathcal{N}}(S)$ and the *hybrid LSA associated phylogenetic diversity* $PD^{LSA_{hyb}}_{\mathcal{N}}(S)$ coincide for all subsets $S \subseteq X$ of taxa.

Example 12. Consider the rooted phylogenetic network $\mathcal{N}_8$ on $X = \{A, B, C, D\}$ depicted in Figure 17. Exemplarily, we set $S = \{A, B\}$ and calculate all versions of the *LSA associated phylogenetic diversity* for S. Note that for $\mathcal{N}_8$ the *LSA tree*, the *hybrid LSA tree* and the *Maximum Likelihood LSA tree* coincide, thus,

$$\begin{aligned} PD^{LSA}_{\mathcal{N}_8}(S) &\coloneqq PD_{\mathcal{T}_{LSA}(\mathcal{N}_8)}(S) = 6, \\ PD^{LSA_{hyb}}_{\mathcal{N}_8}(S) &\coloneqq PD_{\mathcal{T}^{hyb}_{LSA}(\mathcal{N}_8)}(S) = 6, \\ PD^{LSA_{ML}}_{\mathcal{N}_8}(S) &\coloneqq PD_{\mathcal{T}^{ML}_{LSA}(\mathcal{N}_8)}(S) = 6. \end{aligned}$$

Analogously, the different versions of the *LSA associated phylogenetic diversity* can be calculated for all subset $S \subseteq X$ of taxa. Table 1 summarizes the results.

Example 13. Now, consider the rooted phylogenetic network $\mathcal{N}_8'$ on $X = \{A, B, C, D\}$ and its associated *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N}_8')$, *hybrid LSA tree* $\mathcal{T}_{LSA}^{hyb}(\mathcal{N}_8')$ and *Maximum Likelihood LSA tree* $\mathcal{T}_{LSA}^{ML}(\mathcal{N}_8')$ depicted in Figures 18 and 19. Again, we set $S = \{A, B\}$ and calculate all version of the *LSA associated phylogenetic diversity* for S.

$$\begin{aligned} PD_{\mathcal{N}_8'}^{LSA}(S) &\coloneqq PD_{\mathcal{T}_{LSA}(\mathcal{N}_8')}(S) = 3 + \frac{7}{2} = \frac{13}{2}, \\ PD_{\mathcal{N}_8'}^{LSA_{hyb}}(S) &\coloneqq PD_{\mathcal{T}_{LSA}^{hyb}(\mathcal{N}_8')}(S) = 3 + \frac{11}{3} = \frac{20}{3}, \\ PD_{\mathcal{N}_8'}^{LSA_{ML}}(S) &\coloneqq PD_{\mathcal{T}_{LSA}^{ML}(\mathcal{N}_8')}(S) = 3 + 4 = 7. \end{aligned}$$

Thus, in general, the different version of the *LSA associated phylogenetic diversity* do not coincide.

Remark. Above examples suggest that the calculation of any version of the *LSA associated phylogenetic diversity* is easy to accomplish. In order to calculate the *LSA associated phylogenetic diversity* of a subset $S \subseteq X$ of taxa, we simply have to calculate the weight of the sub-arborescence of the *(ordinary/hybrid/Maximum Likelihood) LSA tree* spanning S and the root, which can be achieved in linear time (cf. Hwang et al. [21]).

Recall, however, that the construction of the *(ordinary/hybrid/Maximum Likelihood) LSA tree* prior to the calculation of the *LSA associated phylogenetic diversity* itself may be a limiting factor in practice, because it involves the enumeration of all paths between the reticulation nodes of the network and their *lowest stable ancestors*, respectively (cf. Remark on page 28).

4.4. Inherited Phylogenetic Diversity

Given hybridization probabilities for a phylogenetic network $\mathcal{N}$ on X, we will now propose one last approach of generalizing *phylogenetic diversity* from trees to networks. In this approach we try to infer the *phylogenetic diversity* of a subset $S \subseteq X$ of taxa directly from the network by considering the probability of an edge e contributing to the diversity of the set S, i.e. we consider the probability that the diversity represented by the edge e is preserved or inherited in S.

Definition 33 (Inherited phylogenetic diversity)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on some taxon set X and let $E_T(\mathcal{N})$ be the set of tree edges of $\mathcal{N}$. Now let $S \subseteq X$ be a subset of taxa. For each tree edge $e \in E_T(\mathcal{N})$, we use λ_e to denote the length of e and p_e^S to denote the probability of the diversity represented by e being preserved in S, i.e. p_e^S is the probability of e being included in a smallest arborescence (i.e. an arborescence that does not contain any additional nodes or edges than those required to connect S and the root) spanning S and the root. Then the *inherited phylogenetic diversity* of S is defined as

$$IPD(S) = \sum_{e \in E_T(\mathcal{N})} p_e^S \cdot \lambda_e. \tag{4.16}$$

Remark. Note that an edge $e \in E_T(\mathcal{N})$ can only be included in a smallest arborescence spanning S and the root, if e is contained in at least one path between the root and a taxon in S. Thus, if e is not contained in any such path, the probability p_e^S will be 0.

However, we have not found an efficient way or algorithm to directly calculate the probability p_e^S of an edge e yet.

One way could be to fix an edge e and then consider the set $\mathcal{A}_S^e$ of all smallest arborescences spanning S and the root that contain this edge. Then the sum of probabilities of these arborescences yields the probability of e contributing diversity to S, i.e.

$$p_e^S = \sum_{A \in \mathcal{A}_S^e} \mathbb{P}(A). \tag{4.17}$$

Example 14. Consider the rooted phylogenetic network $\mathcal{N}_8'$ on $X = \{A, B, C, D\}$ depicted in Figure 13. Exemplarily, we calculate the *inherited phylogenetic diversity* of different subsets $S \subseteq X$.

- $S = \{A\}$:
  For $S = \{A\}$, the only tree edges that add diversity to A are $e_1 := (\rho, u)$ and $e_2 := (u, A)$. Both for e_1 and e_2 there is only one arborescence spanning A and

the root, that contains e_1 and e_2, respectively. These arborescences both have probability 1. Thus, $p^S_{e_1} = 1$ and $p^S_{e_2} = 1$ and

$$IPD(\{A\}) = \underbrace{1 \cdot 1}_{(\rho,u)} + \underbrace{1 \cdot 2}_{(u,A)} = 3.$$

- $S = \{B\}$:
  For $S = \{B\}$, there are four tree edges that B may inherit diversity from: $e_1 := (\rho, u), e_2 := (\rho, v), e_3 := (r_2, w)$ and $e_4 := (w, B)$. For e_1 there is one arborescence that contains e_1 and spans B and the root, which has probability $\frac{1}{3}$, thus, $p^S_{e_1} = \frac{1}{3}$. Similarly, we have $p^S_{e_2} = \frac{2}{3}$, $p^S_{e_3} = 1$ and $p^S_{e_4} = 1$. Thus,

$$IPD(\{B\}) = \underbrace{\frac{1}{3} \cdot 1}_{(\rho,u)} + \underbrace{\frac{2}{3} \cdot 2}_{(\rho,v)} + \underbrace{1 \cdot 1}_{(r_2,w)} + \underbrace{1 \cdot 1}_{(w,B)} = \frac{11}{3}.$$

- $S = \{A, B\}$:
  For $S = \{A, B\}$ we have to consider the tree edges $e_1 := (\rho, u)$, $e_2 := (\rho, v)$, $e_3 := (r_2, w)$, $e_4 := (w, B)$ and $e_5 := (u, A)$. There are two smallest arborescences spanning S and the root that contain e_1. One has probability $\frac{1}{3}$, the other $\frac{2}{3}$, thus, $p^S_{e_1} = \frac{1}{3} + \frac{2}{3} = 1$. Analogously, we have $p^S_{e_2} = \frac{2}{3}$, $p^S_{e_3} = 1$, $p^S_{e_4} = 1$ and $p^S_{e_5} = 1$. Thus,

$$IPD(\{A, B\}) = \underbrace{1 \cdot 1}_{(\rho,u)} + \underbrace{\frac{2}{3} \cdot 2}_{(\rho,v)} + \underbrace{1 \cdot 1}_{(r_2,w)} + \underbrace{1 \cdot 1}_{(w,B)} + \underbrace{1 \cdot 2}_{(u,A)} = \frac{19}{3}.$$

Analogously, the *inherited phylogenetic diversity* can be calculated for all other subsets $S \subseteq X$ of taxa.
Note that we have $PND^{hyb}(\{A\}) = 3$, $PND^{hyb}(\{B\}) = \frac{11}{3}$ and $PND^{hyb}(\{A, B\}) = \frac{1}{3} \cdot 5 + \frac{2}{3} \cdot 7 = \frac{19}{3}$, thus, PND and IPD coincide for the subsets $S \subseteq X$ considered.

Remark. The calculation of the probability p^S_e of an edge e contributing diversity to a subset $S \subseteq X$ of taxa via Equation (4.17) and thus, via the set of smallest arborescences spanning S and the root, suggests that the *hybrid phylogenetic net diversity* $PND^{hyb}(S)$ and the *inherited phylogenetic diversity* $IPD(S)$ coincide for all $S \subseteq X$.

Proposition 4. *Let $\mathcal{N}$ be a rooted phylogenetic network on some taxon set X and let $S \subseteq X$ be a subset of taxa. Then we have*

$$PND^{hyb}(S) = IPD(S) \tag{4.18}$$

for all $S \subseteq X$.

Proof. Let $\mathcal{N}$ be a rooted phylogenetic network on X and let $S \subseteq X$ be a subset of taxa. Let $\mathcal{A}_S = \{A_1, \ldots, A_s\}$ be the set of all smallest arborescences spanning S and the root in $\mathcal{N}$ and let $E_T = \{e_1, \ldots, e_t\}$ be the set of tree edges of $\mathcal{N}$. For all $e \in E_T$ let p_e^S denote the probability of the edge e contributing diversity to the set S and let λ_e denote its length. Then we have

$$\begin{aligned} IPD(S) &= \sum_{e \in E_T} p_e^S \cdot \lambda_e \\ &\overset{4.17}{=} \sum_{e \in E_T} \sum_{A \in \mathcal{A}_S^e} \mathbb{P}(A) \cdot \lambda_e \\ &= \sum_{j=1}^{t} \sum_{A \in \mathcal{A}_S^{e_j}} \mathbb{P}(A) \cdot \lambda_{e_j} \\ &= \sum_{A \in \mathcal{A}_S^{e_1}} \mathbb{P}(A) \cdot \lambda_{e_1} + \ldots + \sum_{A \in \mathcal{A}_S^{e_t}} \mathbb{P}(A) \cdot \lambda_{e_t}. \end{aligned}$$

We now factor out $A_1, \ldots, A_s$, respectively, and rearrange the sum in the following way:

$$\begin{aligned} \sum_{A \in \mathcal{A}_S^{e_1}} \mathbb{P}(A) \cdot \lambda_{e_1} + \ldots + \sum_{A \in \mathcal{A}_S^{e_t}} \mathbb{P}(A) \cdot \lambda_{e_t} &= \mathbb{P}(A_1)(\mathbb{1}_{e_1}^{A_1} \lambda_{e_1} + \mathbb{1}_{e_2}^{A_1} \lambda_{e_2} + \ldots \mathbb{1}_{e_t}^{A_1}) + \ldots \\ &\quad \ldots + \mathbb{P}(A_s)(\mathbb{1}_{e_1}^{A_s} \lambda_{e_1} + \mathbb{1}_{e_2}^{A_s} \lambda_{e_2} + \ldots \mathbb{1}_{e_t}^{A_s}) \\ &= \sum_{i=1}^{s} \mathbb{P}(A_i) \Big(\sum_{j=1}^{t} \mathbb{1}_{e_j}^{A_i} \lambda e_j \Big), \end{aligned}$$

where

$$\mathbb{1}_{e_j}^{A_i} = \begin{cases} 1, & \text{if } A_i \text{ contains the edge } e_j; \\ 0, & \text{else.} \end{cases}$$

However,

$$\sum_{j=1}^{t} \mathbb{1}_{e_j}^{A_i} \lambda e_j = weight(A_i)$$

for all arborescences A_i, $i = 1, \ldots, s$, and thus,

$$\begin{aligned} \sum_{i=1}^{s} \mathbb{P}(A_i) \Big(\sum_{j=1}^{t} \mathbb{1}_{e_j}^{A_i} \lambda e_j \Big) &= \sum_{i=1}^{s} \mathbb{P}(A_i) \cdot weight(A_i) \\ &= \sum_{A \in \mathcal{A}_S} \mathbb{P}(A) \cdot weight(A) \\ &= PND^{hyb}(S). \end{aligned}$$

In total we have

$$IPD(S) = PND^{hyb}(S),$$

as claimed. □

4.5. Conclusion

In order to extend the concept of *phylogenetic diversity* from trees to networks, we have developed several approaches, following three main principles: the calculation of spanning arborescences in a phylogenetic network, the computation of the (multi)set of phylogenetic trees displayed by a network and the construction of the *LSA tree* associated with it. For all approaches we have both considered measures that incorporate hybridization probabilities (if given) and measures that are independent of hybridization probabilities.

- Measures of *phylogenetic diversity* based on smallest arborescences:
  - *phylogenetic net diversity* PND,
  - *hybrid phylogenetic net diversity* PND^{hyb},
  - *Maximum Likelihood phylogenetic net diversity* PND^{ML},
  - *inherited phylogenetic diversity* IPD.
- Measures of *phylogenetic diversity* based on the (multi)set of embedded trees:

  embedded phylogenetic diversity $PD^{*}_{\mathsf{T}(\mathcal{N})}$ with

  $$* \in \{\min, \max, \sum, \varnothing\, \varnothing_{hyb}, ML\}.$$

- Measures of *phylogenetic diversity* based on the *LSA tree*:

  LSA associated phylogenetic diversity PD^{LSA}, $PD^{LSA_{hyb}}$, $PD^{LSA_{ML}}$.

The calculation of the *phylogenetic net diversity* $PND(S)$ involves the computation of a minimum cost arborescence spanning S and the root, which is an instance of the NP-hard directed Steiner tree problem (Floudas and Pardalos [13], p. 3731). The computation of $PND(S)$ for all subsets $S \subseteq X$ of taxa may therefore be infeasible in practice. In case of the *hybrid phylogenetic net diversity* $PND^{hyb}(S)$ and the *Maximum Likelihood phylogenetic net diversity* $PND^{ML}(S)$, we do not consider the minimum cost arborescence spanning S and the root, but all smallest arborescences (i.e. arborescences that do not contain any additional edges than those required to span S and the root). This requires the enumeration of all such arborescences, whose

number may be exponential in the number of reticulation nodes. Thus, the calculation of $PND^{hyb}(S)$ and $PND^{ML}(S)$ may be infeasible for phylogenetic networks with a high number of reticulation nodes.
The same problem arises for any measure of the *LSA associated phylogenetic diversity*, which involves the construction of the *(weighted/hybrid/Maximum Likelihood) LSA tree* associated with the network and thus, the enumeration of all paths between all reticulation nodes of the network and their lowest stable ancestors, respectively. Again, this number may be exponential in the number of reticulation nodes. However, the structure of the network, for which we can observe an exponential number of paths between a reticulation node and its lowest stable ancestor (cf. Figure 11) is very 'artificial' and may not appear in biological hybridization networks. Thus, in practice, the construction of the *LSA tree* may still be feasible, which is also supported by tests on random phylogenetic networks (cf. Chapter 6.3).
Last, but not least, all measures of *embedded phylogenetic diversity* depend on the (multi)set of phylogenetic X-trees displayed by a network $\mathcal{N}$ on X, whose determination is, again, an NP-hard problem (cf. Linz et al. [25]). Thus, the calculation of the *embedded phylogenetic diversity* for all subsets $S \subseteq X$ of taxa may be infeasible in practice.

Under computational aspects it may therefore be necessary to develop approximations for the different measures of generalized *phylogenetic diversity.*

Under biologically aspects, on the other hand, it may also be necessary to question the biological plausibility of all measures. Intuitively, we consider the approaches $PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}$ for networks without hybridization probabilities and $PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}$ and $PND^{hyb} = IPD$ for networks with given hybridization probabilities to be the most promising, because they take into account most of the information about the evolutionary relationships between the taxa. Considering the fact that evolution at the nucleotide level is still treelike, motivates in particular the consideration of the (multi)set $\mathsf{T}(\mathcal{N})$ of trees displayed by a network. Even though all measures of *embedded phylogenetic diversity* consider the (multi)set $\mathsf{T}(\mathcal{N})$, only $PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}$ and $PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}$ use all the information of $\mathsf{T}(\mathcal{N})$, while $PD^{*}_{\mathsf{T}(\mathcal{N})}$ with $* \in \{\min, \max, ML\}$ only uses parts of it and discards all other information.
Similarly, all measures of *phylogenetic diversity* based on the *LSA tree* discard information present in the network, because the *LSA tree* reduces the network to its most basic treelike content. However, in practice the calculation of the *LSA tree* and thus, the computation of the *LSA associated phylogenetic diversity* seems to be much more feasible than the calculation of the (multi)set $\mathsf{T}(\mathcal{N})$ and, subsequently,

the computation of the *embedded phylogenetic diversity* (cf. Chapter 6.3). Thus, the consideration of as much information as possible seems to be at the expense of loss of feasibility (in terms of computation times), and vice versa fast computation times and thus, a high practicability, seem to be at the expense of loss of evolutionary information.

Note, however, that the information about the *phylogenetic diversity* of a subset $S \subseteq X$ of taxa in itself is not very useful to taxon prioritization decisions. Thus, we will now turn our attention to generalized biodiversity indices, which, however, to some extent build up on the concept of generalized *phylogenetic diversity*.

5. Generalization of phylogenetic diversity indices to hybridization networks

After proposing different ways of generalizing the concept of *phylogenetic diversity* from trees to networks, we will now turn our attention to biodiversity indices and show how to use them in the context of hybridization networks.

We will focus on the *Fair Proportion Index* and the different versions of the *Shapley Value* introduced in Chapter 2.3, all of which have been suggested as taxon prioritization tools in biodiversity conservation.

Even though all indices are closely related, they differ significantly in their definition and computation. While the calculation of the *Fair Proportion Index* is directly based on a given rooted phylogenetic tree (cf. Definition 5), the definition of the *Shapley Value* is based on the *phylogenetic diversity* of all possible subsets of taxa, and thus, only indirectly on a given (un)rooted phylogenetic tree (cf. Definitions 6, 7, 8). To be precise, the calculation of the different versions of the *Shapley Value* involves two steps:

1. Calculation of the *phylogenetic diversity* for all subsets of taxa based on a given phylogenetic tree.

2. Calculation of the *Shapley Value* for all taxa based on the *phylogenetic diversity* calculated in step 1.

This implies that we have two possibilities when extending the *Shapley Value* from trees to networks: We can either use any generalized definition of *phylogenetic diversity* (e.g. the *phylogenetic net diversity*, the *embedded phylogenetic diversity*, the *LSA associated phylogenetic diversity*, etc.) introduced in Chapter 4 and calculate the *Shapley Value* based on this measure, or we can reduce the network to its treelike content (e.g. via the (multi)set of embedded trees or the *LSA tree*) and calculate the *Shapley Value* based on these trees. In the following, we will analyze and compare both approaches.

We will, however, start with the *Fair Proportion Index*, for which we will also use the second approach, i.e. the reduction to the treelike content of a network, before we try to directly adapt the definition of the *Fair Proportion Index* by considering all paths between the root and a taxon.

5.1. The Fair Proportion Index

Recall that the *Fair Proportion Index* of a taxon $a \in X$, where X is the leaf set of a rooted phylogenetic X-tree, was defined as a weighted sum of edge lengths of edges on

the path from the root to the leaf, where each edge is weighted according to its number of descendent leaves (cf. Definition 5).

In order to use the *Fair Proportion Index* as a prioritization criterion for the taxa of a rooted phylogenetic network $\mathcal{N}$ on X, we consider two approaches that reduce $\mathcal{N}$ to its treelike content, namely the (multi)set of embedded trees (cf. Chapter 3.2) and the *LSA tree* associated with $\mathcal{N}$ (cf. Chapter 3.3).

We then try to adapt the definition of the *Fair Proportion Index* to networks by considering all paths between the root and a taxon.

5.1.1. Embedded Fair Proportion Index

Similar to the different versions of the *embedded phylogenetic diversity*, we now introduce the *embedded Fair Proportion Index* for hybridization networks.

Definition 34 (Embedded Fair Proportion Index)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on some taxon set X and let $\mathsf{T}(\mathcal{N})$ be the (multi)set of all rooted phylogenetic X-trees displayed by $\mathcal{N}$. Then we use $FP^{*}_{\mathsf{T}(\mathcal{N})}(a)$ to denote the *embedded Fair Proportion Index* of a taxon $a \in X$, where $*$ stands for $\min, \max, \sum, \varnothing$ and define

$$FP^{\min}_{\mathsf{T}(\mathcal{N})}(a) := \min_{\mathcal{T}\in\mathsf{T}(\mathcal{N})}\{FP_{\mathcal{T}}(a)\}, \tag{5.1}$$

$$FP^{\max}_{\mathsf{T}(\mathcal{N})}(a) := \max_{\mathcal{T}\in\mathsf{T}(\mathcal{N})}\{FP_{\mathcal{T}}(a)\}, \tag{5.2}$$

$$FP^{\sum}_{\mathsf{T}(\mathcal{N})}(a) := \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} FP_{\mathcal{T}}(a) \text{ and} \tag{5.3}$$

$$FP^{\varnothing}_{\mathsf{T}(\mathcal{N})}(a) := \frac{1}{|\,\mathsf{T}(\mathcal{N})\,|}\sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} FP_{\mathcal{T}}(a), \tag{5.4}$$

$$\tag{5.5}$$

where $|\,\mathsf{T}(\mathcal{N})\,|$ is the number of phylogenetic X-trees displayed by $\mathcal{N}$.
If hybridization probabilities are given for $\mathcal{N}$, we also consider

$$FP^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(a) := \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} \mathbb{P}(\mathcal{T})\cdot FP_{\mathcal{T}}(a) \text{ and} \tag{5.6}$$

$$FP^{ML}_{\mathsf{T}(\mathcal{N})}(a) := FP_{\mathcal{T}^*}(a) \text{ with } \mathcal{T}^* = \operatorname*{argmax}_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} \mathbb{P}(\mathcal{T}), \tag{5.7}$$

where $\mathbb{P}(\mathcal{T})$ is the probability of $\mathcal{T}$ and $\mathcal{T}^*$ is the most likely embedded tree. If the argmax is not unique, we arbitrarily choose one of the embedded trees with maximum probability (cf. Definition 22).

Example 15. Consider the rooted phylogenetic network $\mathcal{N}_8$ on $X = \{A, B, C, D\}$ and the *Fair Proportion Indices* for its embedded trees depicted in Figure 23. Note that we have $\mathbb{P}(\mathcal{T}_1') = \frac{1}{6}$, $\mathbb{P}(\mathcal{T}_2') = \frac{1}{2}$, $\mathbb{P}(\mathcal{T}_3') = \frac{1}{4}$ and $\mathbb{P}(\mathcal{T}_4') = \frac{1}{12}$. Thus, $\mathcal{T}_2'$ is the most likely embedded tree.
Exemplarily, we now calculate all versions of the *embedded Fair Proportion Index* for the taxon $A \in X$.

$$\begin{aligned} FP^{\min}_{\mathsf{T}(\mathcal{N}_8)}(A) &= \min_{\mathcal{T}\in\mathsf{T}(\mathcal{N}_8)} FP_{\mathcal{T}}(A) \\ &= \min\{FP_{\mathcal{T}_1'}(A), FP_{\mathcal{T}_2'}(A), FP_{\mathcal{T}_3'}(A), FP_{\mathcal{T}_4'}(A)\} \\ &= \min\left\{3, 3, \frac{7}{3}, \frac{5}{2}\right\} \\ &= \frac{7}{3}, \end{aligned}$$

$$\begin{aligned} FP^{\max}_{\mathsf{T}(\mathcal{N}_8)}(A) &= \max_{\mathcal{T}\in\mathsf{T}(\mathcal{N}_8)} FP_{\mathcal{T}}(A) \\ &= \max\{FP_{\mathcal{T}_1'}(A), FP_{\mathcal{T}_2'}(A), FP_{\mathcal{T}_3'}(A), FP_{\mathcal{T}_4'}(A)\} \\ &= \max\left\{3, 3, \frac{7}{3}, \frac{5}{2}\right\} \\ &= 3, \end{aligned}$$

$$\begin{aligned} FP^{\varnothing}_{\mathsf{T}(\mathcal{N}_8)}(A) &= \frac{1}{|\mathsf{T}(\mathcal{N}_8)|} \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N}_8)} FP_{\mathcal{T}}(A) \\ &= \frac{1}{4}\Big(FP_{\mathcal{T}_1'}(A) + FP_{\mathcal{T}_2'}(A) + FP_{\mathcal{T}_3'}(A) + FP_{\mathcal{T}_4'}(A)\Big) \\ &= \frac{1}{4}\Big(3 + 3 + \frac{7}{3} + \frac{5}{2}\Big) \\ &= \frac{65}{24}, \end{aligned}$$

$$\begin{aligned} FP^{\Sigma}_{\mathsf{T}(\mathcal{N}_8)}(A) &= \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N}_8)} FP_{\mathcal{T}}(A) \\ &= FP_{\mathcal{T}_1'}(A) + FP_{\mathcal{T}_2'}(A) + FP_{\mathcal{T}_3'}(A) + FP_{\mathcal{T}_4'}(A) \\ &= 3 + 3 + \frac{7}{3} + \frac{5}{2} \\ &= \frac{65}{6}, \end{aligned}$$

$$
\begin{aligned}
FP^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_8)}(A) &= \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N}_8)} \mathbb{P}(\mathcal{T})\cdot FP_{\mathcal{T}}(A) \\
&= \mathbb{P}(\mathcal{T}_1')\cdot FP_{\mathcal{T}_1'}(A) + \mathbb{P}(\mathcal{T}_2')\cdot FP_{\mathcal{T}_2'}(A) + \mathbb{P}(\mathcal{T}_3')\cdot FP_{\mathcal{T}_3'}(A) + \mathbb{P}(\mathcal{T}_4')\cdot FP_{\mathcal{T}_4'}(A) \\
&= \frac{1}{6}\cdot 3 + \frac{1}{2}\cdot 3 + \frac{1}{4}\cdot\frac{7}{3} + \frac{1}{12}\cdot\frac{5}{2} \\
&= \frac{67}{24},
\end{aligned}
$$

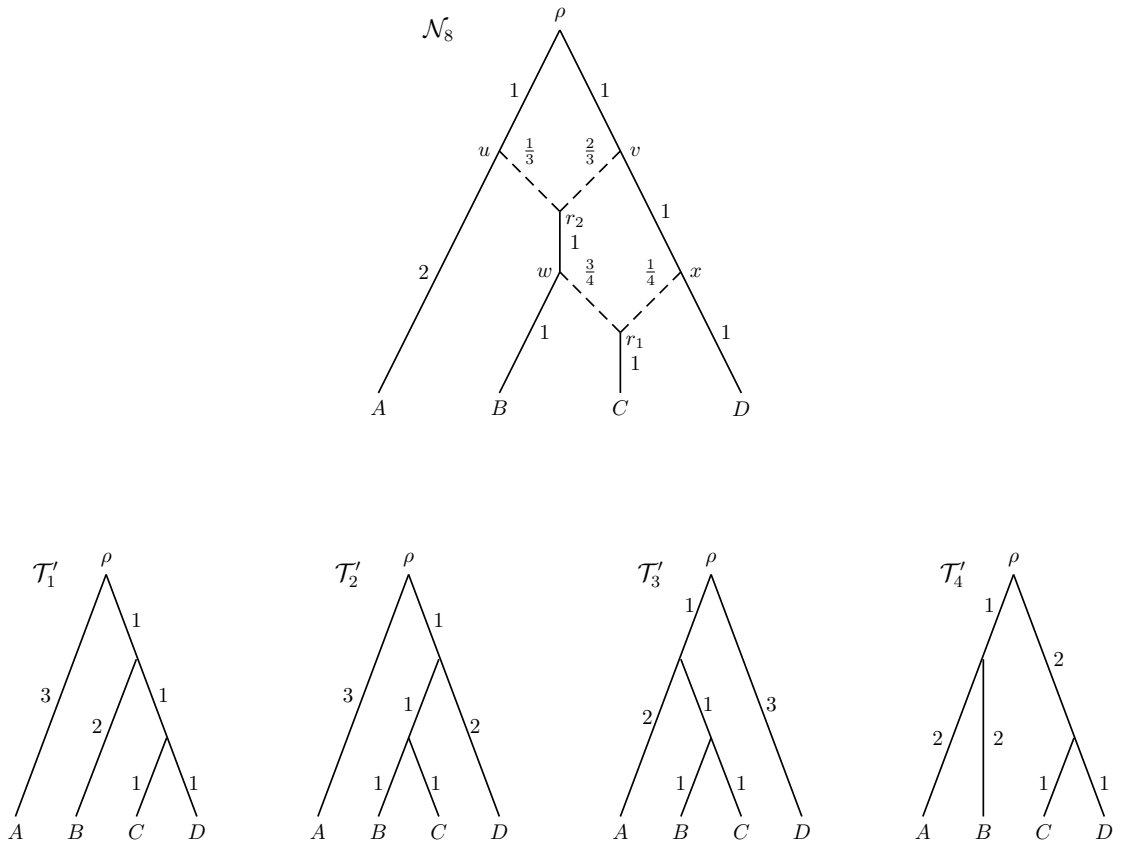
and

$$
\begin{aligned}
FP^{ML}_{\mathsf{T}(\mathcal{N}_8)}(A) &= FP_{\mathcal{T}_2'}(A) \\
&= 3.
\end{aligned}
$$

Analogously, the *embedded Fair Proportion Indices* can be calculated for all other taxa in X. Table 2 summarizes the results.

Table 2: *Embedded Fair Proportion Indices* for the rooted phylogenetic network $\mathcal{N}_8$

$a \in X$	$FP^{\min}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$FP^{\max}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$FP^{\sum}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$FP^{\varnothing}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$FP^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$FP^{ML}_{\mathsf{T}(\mathcal{N}_8)}(a)$
A	$\frac{7}{3}$	3	$\frac{65}{6}$	$\frac{65}{24}$	$\frac{67}{24}$	3
B	$\frac{11}{6}$	$\frac{5}{2}$	$\frac{17}{2}$	$\frac{17}{8}$	$\frac{71}{36}$	$\frac{11}{6}$
C	$\frac{11}{6}$	2	$\frac{15}{2}$	$\frac{15}{8}$	$\frac{133}{72}$	$\frac{11}{6}$
D	$\frac{11}{6}$	3	$\frac{55}{6}$	$\frac{55}{24}$	$\frac{43}{18}$	$\frac{7}{3}$
$\sum$	$\frac{47}{6}$	$\frac{21}{2}$	36	9	9	9



$FP_{\mathcal{T}_1'}(A) = SV_{\mathcal{T}_1'}(A) = 3$	$FP_{\mathcal{T}_2'}(A) = SV_{\mathcal{T}_2'}(A) = 3$	$FP_{\mathcal{T}_3'}(A) = SV_{\mathcal{T}_3'}(A) = \frac{7}{3}$	$FP_{\mathcal{T}_4'}(A) = SV_{\mathcal{T}_4'}(A) = \frac{5}{2}$
$FP_{\mathcal{T}_1'}(B) = SV_{\mathcal{T}_1'}(B) = \frac{7}{3}$	$FP_{\mathcal{T}_2'}(B) = SV_{\mathcal{T}_2'}(B) = \frac{11}{6}$	$FP_{\mathcal{T}_3'}(B) = SV_{\mathcal{T}_3'}(B) = \frac{11}{6}$	$FP_{\mathcal{T}_4'}(B) = SV_{\mathcal{T}_4'}(B) = \frac{5}{2}$
$FP_{\mathcal{T}_1'}(C) = SV_{\mathcal{T}_1'}(C) = \frac{11}{6}$	$FP_{\mathcal{T}_2'}(C) = SV_{\mathcal{T}_2'}(C) = \frac{11}{6}$	$FP_{\mathcal{T}_3'}(C) = SV_{\mathcal{T}_3'}(C) = \frac{11}{6}$	$FP_{\mathcal{T}_4'}(C) = SV_{\mathcal{T}_4'}(C) = 2$
$FP_{\mathcal{T}_1'}(D) = SV_{\mathcal{T}_1'}(D) = \frac{11}{6}$	$FP_{\mathcal{T}_2'}(D) = SV_{\mathcal{T}_2'}(D) = \frac{7}{3}$	$FP_{\mathcal{T}_3'}(D) = SV_{\mathcal{T}_3'}(D) = 3$	$FP_{\mathcal{T}_4'}(D) = SV_{\mathcal{T}_4'}(D) = 2$

Fig. 23: Rooted phylogenetic network $\mathcal{N}_8$ on $X = \{A, B, C, D\}$ with the *Fair Proportion Indices* and *original Shapley Values* for its embedded trees.

Remarks.

- If all embedded trees $\mathcal{T} \in \mathsf{T}(\mathcal{N})$ are equally likely, $FP^{\varnothing}_{\mathsf{T}(\mathcal{N})}(a)$ and $FP^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(a)$ coincide for all taxa $a \in X$.

- Note that $FP^{\min}_{\mathsf{T}(\mathcal{N})}$ and $FP^{\max}_{\mathsf{T}(\mathcal{N})}$ are not *efficient* (cf. page 8), i.e.

$$\sum_{a \in X} FP^{\min(\max)}_{\mathsf{T}(\mathcal{N})}(a) \neq length(\mathcal{N}),$$

  where $length(\mathcal{N})$ is the sum of branch lengths in $\mathcal{N}$.
  For $FP^{\sum}_{\mathsf{T}(\mathcal{N})}$ we have

$$\sum_{a \in X} FP^{\sum}_{\mathsf{T}(\mathcal{N})}(a) = |\,\mathsf{T}(\mathcal{N})\,| \cdot length(\mathcal{N}).$$

- Even though the computation of the *Fair Proportion Index* for a phylogenetic tree is easy to accomplish, the computation of the *embedded Fair Proportion Index* is not. This is due to the fact, that the computation of the *embedded Fair Proportion Index* requires the (multi)set of embedded trees $\mathsf{T}(\mathcal{N})$, whose determination is an NP-hard problem (cf. Linz et al. [25]).

5.1.2. LSA associated Fair Proportion Index

We will now use the *LSA tree*, the *hybrid LSA tree* and the *Maximum Likelihood LSA tree* (cf. Definition 18) associated with a phylogenetic network $\mathcal{N}$ in order to compute the *Fair Proportion Index* for the taxa of $\mathcal{N}$.

Definition 35 (LSA associated Fair Proportion Index)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on some taxon set X. Let $a \in X$ be a taxon in X. Then we define

$$FP^{LSA}(a) := FP_{\mathcal{T}_{LSA}(\mathcal{N})}(a), \tag{5.8}$$

where $FP_{\mathcal{T}_{LSA}(\mathcal{N})}(a)$ is the *Fair Proportion Index* of a in the *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N})$ associated with $\mathcal{N}$. If hybridization probabilities are given for $\mathcal{N}$, we also consider

$$FP^{LSA_{hyb}}(a) := FP_{\mathcal{T}^{hyb}_{LSA}(\mathcal{N})}(a) \text{ and} \tag{5.9}$$

$$FP^{LSA_{ML}}(a) := FP_{\mathcal{T}^{ML}_{LSA}(\mathcal{N})}(a), \tag{5.10}$$

where $FP_{\mathcal{T}^{hyb}_{LSA}(\mathcal{N})}(a)$ is the *Fair Proportion Index* of a in the *hybrid LSA tree* $\mathcal{T}^{hyb}_{LSA}(\mathcal{N})$ associated with $\mathcal{N}$ and $FP_{\mathcal{T}^{ML}_{LSA}(\mathcal{N})}(a)$ is the *Fair Proportion Index* of a in the *Maximum*

Likelihood LSA tree $\mathcal{T}_{LSA}^{ML}(\mathcal{N})$ associated with $\mathcal{N}$.
We will call all three versions *LSA associated Fair Proportion Index* and use the superscript to indicate which version of the *LSA tree* was used.

Example 16. Consider the rooted phylogenetic network $\mathcal{N}_8$ on $X = \{A, B, C, D\}$ and its *hybrid LSA tree* depicted in Figure 17. For $\mathcal{N}_8$ all versions of the *LSA tree* coincide, i.e.

$$\mathcal{T}_{LSA}(\mathcal{N}_8) = \mathcal{T}_{LSA}^{hyb}(\mathcal{N}_8) = \mathcal{T}_{LSA}^{ML}(\mathcal{N}).$$

Thus, we have

$$\begin{aligned} FP^*(A) &= 3, \\ FP^*(B) &= 3, \\ FP^*(C) &= 3, \\ FP^*(D) &= 3, \end{aligned}$$

where $* \in \{LSA, LSA_{hyb}, LSA_{ML}\}$.
Note that the treeshape of any *LSA tree* associated with a network $\mathcal{N}$ reduces the networks to its most fundamental treelike content. In case of $\mathcal{N}_8$ this results in a 'rooted startree', which only carries the information that all taxa descend from the root, but does not contain any information about the evolutionary relationships among the taxa. In particular, all taxa receive identical *LSA associated Fair Proportion Indices*, which hinders taxon prioritization decisions.

5.1.3. The Net Fair Proportion Index

After reducing a network $\mathcal{N}$ on X to its treelike content in order to calculate the *Fair Proportion Index* for its taxa, we now try to directly adapt the definition of the *Fair Proportion Index* to networks by considering all paths between the root and a taxon.

W.l.o.g. we assume the network $\mathcal{N}$ to come with hybridization probabilities (if no hybridization probabilities are given for $\mathcal{N}$, we set the probability $\gamma_e = \frac{1}{2}$ for all reticulation edges).

The idea is now to define the *Net Fair Proportion Index* of a taxon $a \in X$ by considering all paths from the root to a and calculating a value for each path individually. Similar to the original *Fair Proportion Index*, we calculate this value as a weighted sum of branch lengths, where each branch length is weighted according to the number of its descendants. However, we additionally weight the possible descendants of an edge e by their probability of actually being a descendant of this edge. We then use the

weighted mean of these values for all paths, where a path is weighted according to its probability, and call the resulting value the *Net Fair Proportion Index*.

Definition 36 (Net Fair Proportion Index)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on some taxon set X. Let λ_e denote the length of an edge e in $\mathcal{N}$ and let D_e denote the set of leaves that are descendants of e.
For each leaf $d \in D_e$ we use $\mathbb{P}^e_{desc}(d)$ to denote the probability of d being descendent from e and calculate $\mathbb{P}^e_{desc}(d)$ as

$$\mathbb{P}^e_{desc}(d) = \sum_{P \in \mathcal{P}_{e,d}} \mathbb{P}(P), \tag{5.11}$$

where $\mathcal{P}_{e,d}$ is the set of paths from the endpoint of e to the leaf d in $\mathcal{N}$ and $\mathbb{P}(P)$ is the probability of any such path (cf. Definition 23 for the probability of a path).
Now let $a \in X$ be a taxon of $\mathcal{N}$ and let $\mathcal{P}_{\rho a}$ be the set of all paths from ρ to a in $\mathcal{N}$. Then we define the *Net Fair Proportion Index* of a as

$$NFP(a) = \sum_{P \in \mathcal{P}_{\rho a}} \mathbb{P}(P) \cdot \Big(\sum_{e \in P} \frac{\lambda_e}{\sum\limits_{d \in D_e} \mathbb{P}^e_{desc}(d)} \Big). \tag{5.12}$$

Example 17. Consider the rooted phylogenetic network $\mathcal{N}_8$ on $X = \{A, B, C, D\}$ depicted in Figure 13. We now calculate the *Net Fair Proportion Index* for taxon $B \in X$:
There are two paths from the root ρ to B in $\mathcal{N}$, namely

$$P_1 = \big((\rho, u), (u, r_2), (r_2, w), (w, B)\big) \text{ with probability } \mathbb{P}(P_1) = \frac{1}{3} \text{ and}$$
$$P_2 = \big((\rho, v), (v, r_2), (r_2, w), (w, B)\big) \text{ with probability } \mathbb{P}(P_2) = \frac{2}{3}.$$

Consider, for example, the edge $e = (\rho, u)$. The set of possible descendants from e consists of the taxa A, B and C, thus, $D_e = \{A, B, C\}$. The probabilities of these taxa descending from e calculate as

$$\mathbb{P}^e_{desc}(A) = 1,$$
$$\mathbb{P}^e_{desc}(B) = \frac{1}{3} \cdot 1 \cdot 1 = \frac{1}{3} \text{ and}$$
$$\mathbb{P}^e_{desc}(C) = \frac{1}{3} \cdot 1 \cdot \frac{3}{4} \cdot 1 = \frac{1}{4}.$$

Analogously, these probabilities can be calculated for all other edges on P_1 and P_2.

Omitting edges of length 0 (i.e. hybridization edges) in the sum, we have

$$
\begin{aligned}
NFP(B) &= \frac{1}{3}\Bigg(\underbrace{\frac{1}{1}}_{(w,B)} + \underbrace{\frac{1}{\underbrace{1}_{B} + \underbrace{\frac{3}{4}}_{C}}}_{(r_2,w)} + \underbrace{\frac{1}{\underbrace{1}_{A} + \underbrace{\frac{1}{3}}_{B} + \underbrace{\frac{1}{4}}_{C}}}_{(\rho,u)}\Bigg) \\
&\quad + \frac{2}{3}\Bigg(\underbrace{\frac{1}{1}}_{(w,B)} + \underbrace{\frac{1}{\underbrace{1}_{B} + \underbrace{\frac{3}{4}}_{C}}}_{(r_2,w)} + \underbrace{\frac{1}{\underbrace{1}_{D} + \underbrace{\frac{2}{3}}_{B} + \underbrace{\frac{1}{4} + \frac{2}{3}\cdot\frac{3}{4}}_{C}}}_{(\rho,v)}\Bigg) \\
&= \frac{1}{3}\cdot\frac{293}{133} + \frac{2}{3}\cdot\frac{403}{203} \\
&= \frac{23811}{11571} \\
&\approx 2.06.
\end{aligned}
$$

Similar calculations yield

$$
\begin{aligned}
NFP(A) &= \frac{50}{19} \approx 2.63, \\
NFP(C) &= \frac{40437}{19285} \approx 2.10 \text{ and} \\
NFP(D) &= \frac{321}{145} \approx 2.21.
\end{aligned}
$$

Note that

$$
\begin{aligned}
\sum_{a\in X} NFP(a) &= \frac{50}{19} + \frac{23811}{11571} + \frac{40437}{19285} + \frac{321}{145} \\
&= 9,
\end{aligned}
$$

thus, the sum of the *Net Fair Proportion Indices* equals the sum of edge lengths in $\mathcal{N}_8$.

Remarks.

- By definition of the *Net Fair Proportion Index*, this measure is *efficient*, i.e.

  $$\sum_{a\in X} NFP(a) = length(\mathcal{N})$$

  for a rooted phylogenetic network $\mathcal{N}$ on X.

- Example 17 shows that the calculation of the *Net Fair Proportion Index* on a

phylogenetic network is much more involved than the calculation of the *Fair Proportion Index* on a phylogenetic tree.

- In order to calculate the NFP for all taxa, we have to consider all possible paths between the root and any of the leaves. In the worst case, the number of paths between the root and a leaf may be exponential in the number of reticulation nodes. Consider, for example, the rooted phylogenetic network $\mathcal{N}_1^*$ depicted in Figure 11 and fix taxon E. As E is a direct descendant of the reticulation node r_5, whose lowest stable ancestor is the root ρ, there are also $2^{r(\mathcal{N}_1^*)-1} + 1 = 17$ paths from ρ to E.

5.2. The Shapley Value

We now turn our attention to the different versions of the *Shapley Value* and show how to calculate these indices in the context of hybridization networks.

Recall that the calculation of any version of the *Shapley Value* was based on the *phylogenetic diversity* of subsets of taxa (cf. Definitions 6, 7, 8), and thus, only indirectly on a given phylogenetic tree.

We will therefore consider two main approaches when generalizing the *Shapley Value* from trees to networks: In a first approach we will reduce the network to its treelike content and introduce the *embedded Shapley Value* and the *LSA associated Shapley Value*. Secondly, we will use the different definitions of generalized *phylogenetic diversity* introduced in Chapter 4 and calculate a value based on these measures, which we will call the *generalized Shapley Value*.

5.2.1. Embedded Shapley Value

Similar to the *embedded Fair Proportion Index*, we now introduce the *embedded Shapley Value*.

Definition 37 (Embedded Shapley Value)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on some taxon set X and let $\mathsf{T}(\mathcal{N})$ be the (multi)set of all rooted phylogenetic X-trees displayed by $\mathcal{N}$. Then we use $SV^*_{\mathsf{T}(\mathcal{N})}(a)$ to denote the *embedded original Shapley Value*

of a taxon $a \in X$, where $*$ stands for $\min, \max, \varnothing, \sum$ and define

$$SV^{\min}_{\mathsf{T}(\mathcal{N})}(a) := \min_{\mathcal{T}\in\mathsf{T}(\mathcal{N})}\{SV_{\mathcal{T}}(a)\}, \tag{5.13}$$

$$SV^{\max}_{\mathsf{T}(\mathcal{N})}(a) := \max_{\mathcal{T}\in\mathsf{T}(\mathcal{N})}\{SV_{\mathcal{T}}(a)\}, \tag{5.14}$$

$$SV^{\varnothing}_{\mathsf{T}(\mathcal{N})}(a) := \frac{1}{|\,\mathsf{T}(\mathcal{N})\,|}\sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} SV_{\mathcal{T}}(a) \text{ and} \tag{5.15}$$

$$SV^{\sum}_{\mathsf{T}(\mathcal{N})}(a) := \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} SV_{\mathcal{T}}(a), \tag{5.16}$$

where $|\,\mathsf{T}(\mathcal{N})\,|$ is the number of phylogenetic X-trees displayed by $\mathcal{N}$.
If hybridization probabilities are given for $\mathcal{N}$, we also consider

$$SV^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(a) := \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} \mathbb{P}(\mathcal{T})\cdot SV_{\mathcal{T}}(a) \text{ and} \tag{5.17}$$

$$SV^{ML}_{\mathsf{T}(\mathcal{N})}(a) := SV_{\mathcal{T}^*}(a) \text{ with } \mathcal{T}^* = \operatorname*{argmax}_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} \mathbb{P}(\mathcal{T}), \tag{5.18}$$

where $\mathbb{P}(\mathcal{T})$ is the probability of $\mathcal{T}$ and $\mathcal{T}^*$ is the most likely embedded tree. If the argmax is not unique, we arbitrarily choose one of the embedded trees with maximum probability (cf. Definition 22).
Analogously, we define the *embedded modified Shapley Value* $\widetilde{SV}^*_{\mathsf{T}(\mathcal{N})}$ and the *embedded unrooted rooted Shapley Value* $\widehat{SV}^*_{\mathsf{T}(\mathcal{N})}$ with $* \in \{\min, \max, \varnothing, \sum, \varnothing_{hyb}, ML\}$.

Example 18. Consider the rooted phylogenetic network $\mathcal{N}_8$ on $X = \{A, B, C, D\}$ and its embedded phylogenetic X-trees depicted in Figure 23. Table 3 contains the different versions of the *Shapley Value* for all taxa and all embedded trees.

Similar to the calculation of the *embedded Fair Proportion Index* (cf. Example 15), we now calculate the different versions of the *embedded Shapley Value* for taxon $A \in X$:

- *Original Shapley Value:*

$$\begin{aligned}
SV^{\min}_{\mathsf{T}(\mathcal{N}_8)}(A) &= \frac{7}{3}\\
SV^{\max}_{\mathsf{T}(\mathcal{N}_8)}(A) &= 3\\
SV^{\varnothing}_{\mathsf{T}(\mathcal{N}_8)}(A) &= \frac{65}{24}\\
SV^{\sum}_{\mathsf{T}(\mathcal{N}_8)}(A) &= \frac{65}{6}\\
SV^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_8)}(A) &= \frac{67}{24}\\
SV^{ML}_{\mathsf{T}(\mathcal{N}_8)}(A) &= 3.
\end{aligned}$$

Table 3: Different versions of the *Shapley Value* for the set of embedded trees in $\mathcal{N}_8$

$\mathcal{T} \in \mathsf{T}(\mathcal{N}_8)$	$\mathcal{T}_1'$	$\mathcal{T}_2'$	$\mathcal{T}_3'$	$\mathcal{T}_4'$
$\mathbb{P}(\mathcal{T})$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{12}$
original SV	$SV_{\mathcal{T}_1'}(A) = 3$	$SV_{\mathcal{T}_2'}(A) = 3$	$SV_{\mathcal{T}_3'}(A) = \frac{7}{3}$	$SV_{\mathcal{T}_4'}(A) = \frac{5}{2}$
	$SV_{\mathcal{T}_1'}(B) = \frac{7}{3}$	$SV_{\mathcal{T}_2'}(B) = \frac{11}{6}$	$SV_{\mathcal{T}_3'}(B) = \frac{11}{6}$	$SV_{\mathcal{T}_4'}(B) = \frac{5}{2}$
	$SV_{\mathcal{T}_1'}(C) = \frac{11}{6}$	$SV_{\mathcal{T}_2'}(C) = \frac{11}{6}$	$SV_{\mathcal{T}_3'}(C) = \frac{11}{6}$	$SV_{\mathcal{T}_4'}(C) = 2$
	$SV_{\mathcal{T}_1'}(D) = \frac{11}{6}$	$SV_{\mathcal{T}_2'}(D) = \frac{7}{3}$	$SV_{\mathcal{T}_3'}(D) = 3$	$SV_{\mathcal{T}_4'}(D) = 2$
modified SV	$\widetilde{SV}_{\mathcal{T}_1'}(A) = \frac{9}{4}$	$\widetilde{SV}_{\mathcal{T}_2'}(A) = \frac{9}{4}$	$\widetilde{SV}_{\mathcal{T}_3'}(A) = \frac{19}{12}$	$\widetilde{SV}_{\mathcal{T}_4'}(A) = \frac{7}{4}$
	$\widetilde{SV}_{\mathcal{T}_1'}(B) = \frac{19}{12}$	$\widetilde{SV}_{\mathcal{T}_2'}(B) = \frac{13}{12}$	$\widetilde{SV}_{\mathcal{T}_3'}(B) = \frac{13}{12}$	$\widetilde{SV}_{\mathcal{T}_4'}(B) = \frac{7}{4}$
	$\widetilde{SV}_{\mathcal{T}_1'}(C) = \frac{13}{12}$	$\widetilde{SV}_{\mathcal{T}_2'}(C) = \frac{13}{12}$	$\widetilde{SV}_{\mathcal{T}_3'}(C) = \frac{13}{12}$	$\widetilde{SV}_{\mathcal{T}_4'}(C) = \frac{5}{4}$
	$\widetilde{SV}_{\mathcal{T}_1'}(D) = \frac{13}{12}$	$\widetilde{SV}_{\mathcal{T}_2'}(D) = \frac{19}{12}$	$\widetilde{SV}_{\mathcal{T}_3'}(D) = \frac{9}{4}$	$\widetilde{SV}_{\mathcal{T}_4'}(D) = \frac{5}{4}$
unrooted rooted SV	$\widehat{SV}_{\mathcal{T}_1'}(A) = \frac{43}{12}$	$\widehat{SV}_{\mathcal{T}_2'}(A) = \frac{43}{12}$	$\widehat{SV}_{\mathcal{T}_3'}(A) = \frac{9}{4}$	$\widehat{SV}_{\mathcal{T}_4'}(A) = \frac{31}{12}$
	$\widehat{SV}_{\mathcal{T}_1'}(B) = \frac{9}{4}$	$\widehat{SV}_{\mathcal{T}_2'}(B) = \frac{19}{12}$	$\widehat{SV}_{\mathcal{T}_3'}(B) = \frac{19}{12}$	$\widehat{SV}_{\mathcal{T}_4'}(B) = \frac{31}{12}$
	$\widehat{SV}_{\mathcal{T}_1'}(C) = \frac{19}{12}$	$\widehat{SV}_{\mathcal{T}_2'}(C) = \frac{19}{12}$	$\widehat{SV}_{\mathcal{T}_3'}(C) = \frac{19}{12}$	$\widehat{SV}_{\mathcal{T}_4'}(C) = \frac{23}{12}$
	$\widehat{SV}_{\mathcal{T}_1'}(D) = \frac{19}{12}$	$\widehat{SV}_{\mathcal{T}_2'}(D) = \frac{9}{4}$	$\widehat{SV}_{\mathcal{T}_3'}(D) = \frac{43}{12}$	$\widehat{SV}_{\mathcal{T}_4'}(D) = \frac{23}{12}$

- *Modified Shapley Value:*

$$\widetilde{SV}^{\min}_{\mathsf{T}(\mathcal{N}_8)}(A) = \frac{19}{12}$$
$$\widetilde{SV}^{\max}_{\mathsf{T}(\mathcal{N}_8)}(A) = \frac{9}{4}$$
$$\widetilde{SV}^{\varnothing}_{\mathsf{T}(\mathcal{N}_8)}(A) = \frac{47}{24}$$
$$\widetilde{SV}^{\Sigma}_{\mathsf{T}(\mathcal{N}_8)}(A) = \frac{47}{6}$$
$$\widetilde{SV}^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_8)}(A) = \frac{49}{24}$$
$$\widetilde{SV}^{ML}_{\mathsf{T}(\mathcal{N}_8)}(A) = \frac{9}{4}.$$

- *Unrooted rooted Shapley Value:*

$$\widehat{SV}^{\min}_{\mathsf{T}(\mathcal{N}_8)}(A) = \frac{9}{4}$$
$$\widehat{SV}^{\max}_{\mathsf{T}(\mathcal{N}_8)}(A) = \frac{43}{12}$$
$$\widehat{SV}^{\varnothing}_{\mathsf{T}(\mathcal{N}_8)}(A) = 3$$
$$\widehat{SV}^{\sum}_{\mathsf{T}(\mathcal{N}_8)}(A) = 12$$
$$\widehat{SV}^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_8)}(A) = \frac{19}{6}$$
$$\widehat{SV}^{ML}_{\mathsf{T}(\mathcal{N}_8)}(A) = \frac{43}{12}.$$

Analogously, the different version of the *embedded Shapley Value* can be calculated for all other taxa in X. Tables 4 – 6 summarize the results.

Table 4: *Embedded original Shapley Values* for the rooted phylogenetic network $\mathcal{N}_8$

$a \in X$	$SV^{\min}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$SV^{\max}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$SV^{\sum}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$SV^{\varnothing}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$SV^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$SV^{ML}_{\mathsf{T}(\mathcal{N}_8)}(a)$
A	$\frac{7}{3}$	3	$\frac{65}{6}$	$\frac{65}{24}$	$\frac{67}{24}$	3
B	$\frac{11}{6}$	$\frac{5}{2}$	$\frac{17}{2}$	$\frac{17}{8}$	$\frac{71}{36}$	$\frac{11}{6}$
C	$\frac{11}{6}$	2	$\frac{15}{2}$	$\frac{15}{8}$	$\frac{133}{72}$	$\frac{11}{6}$
D	$\frac{11}{6}$	3	$\frac{55}{6}$	$\frac{55}{24}$	$\frac{43}{18}$	$\frac{7}{3}$
$\sum$	$\frac{47}{6}$	$\frac{21}{2}$	36	9	9	9

Table 5: *Embedded modified Shapley Values* for the rooted phylogenetic network $\mathcal{N}_8$

$a \in X$	$\widetilde{SV}^{\min}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$\widetilde{SV}^{\max}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$\widetilde{SV}^{\sum}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$\widetilde{SV}^{\varnothing}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$\widetilde{SV}^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$\widetilde{SV}^{ML}_{\mathsf{T}(\mathcal{N}_8)}(a)$
A	$\frac{19}{12}$	$\frac{9}{4}$	$\frac{47}{6}$	$\frac{47}{24}$	$\frac{49}{24}$	$\frac{9}{4}$
B	$\frac{13}{12}$	$\frac{7}{4}$	$\frac{11}{2}$	$\frac{11}{8}$	$\frac{11}{9}$	$\frac{13}{12}$
C	$\frac{13}{12}$	$\frac{5}{4}$	$\frac{9}{2}$	$\frac{9}{8}$	$\frac{79}{72}$	$\frac{13}{12}$
D	$\frac{13}{12}$	$\frac{9}{4}$	$\frac{37}{6}$	$\frac{37}{24}$	$\frac{59}{36}$	$\frac{19}{12}$
$\sum$	$\frac{29}{6}$	$\frac{15}{2}$	24	6	6	6

Table 6: *Embedded unrooted rooted Shapley Values* for the rooted phylogenetic network $\mathcal{N}_8$

$a \in X$	$\widehat{SV}^{\min}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$\widehat{SV}^{\max}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$\widehat{SV}^{\sum}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$\widehat{SV}^{\varnothing}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$\widehat{SV}^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_8)}(a)$	$\widehat{SV}^{ML}_{\mathsf{T}(\mathcal{N}_8)}(a)$
A	$\frac{9}{4}$	$\frac{43}{12}$	12	3	$\frac{19}{6}$	$\frac{43}{12}$
B	$\frac{19}{12}$	$\frac{31}{12}$	8	2	$\frac{16}{9}$	$\frac{19}{12}$
C	$\frac{19}{12}$	$\frac{23}{12}$	$\frac{20}{3}$	$\frac{5}{3}$	$\frac{29}{18}$	$\frac{19}{12}$
D	$\frac{19}{12}$	$\frac{43}{12}$	$\frac{28}{3}$	$\frac{7}{3}$	$\frac{22}{9}$	$\frac{9}{4}$
$\sum$	7	$\frac{35}{3}$	36	9	9	9

Remarks.

- If all embedded trees $\mathcal{T} \in \mathsf{T}(\mathcal{N})$ are equally likely, $SV^{\varnothing}_{\mathsf{T}(\mathcal{N})}(a)$ and $SV^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(a)$ coincide for all taxa $a \in X$. The same holds for $\widetilde{SV}^{\varnothing}_{\mathsf{T}(\mathcal{N})}(a)$ and $\widetilde{SV}^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(a)$ as well as for $\widehat{SV}^{\varnothing}_{\mathsf{T}(\mathcal{N})}(a)$ and $\widehat{SV}^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(a)$.

- Due to the fact that the *modified Shapley Value* is not efficient (cf. Remark on page 11), neither version of the *embedded modified Shapley Value* is efficient.

- In case of the *embedded original Shapley Value*, $SV^{\min}_{\mathsf{T}(\mathcal{N})}$ and $SV^{\max}_{\mathsf{T}(\mathcal{N})}$ are not efficient, while $SV^{\varnothing}_{\mathsf{T}(\mathcal{N})}$, $SV^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}$ and $SV^{ML}_{\mathsf{T}(\mathcal{N})}$ are. For $SV^{\sum}_{\mathsf{T}(\mathcal{N})}$ we have

$$\sum_{a \in X} SV^{\sum}_{\mathsf{T}(\mathcal{N})}(a) = |\,\mathsf{T}(\mathcal{N})\,| \cdot length(\mathcal{N}).$$

  The same properties hold for the *embedded unrooted rooted Shapley Value*.

- The *embedded Fair Proportion Indices* and *embedded Shapley Values* of a rooted phylogenetic network are closely related, which follows from their relatedness on phylogenetic trees.

Proposition 5. *Let $\mathcal{N}$ be a rooted phylogenetic network on some taxon set X with $|X| = n$ and let $\mathsf{T}(\mathcal{N})$ be the (multi)set of all rooted phylogenetic X-trees displayed by $\mathcal{N}$. Let $a \in X$ be a taxon in X. Then we have*

1. $FP^{*}_{\mathsf{T}(\mathcal{N})}(a) = SV^{*}_{\mathsf{T}(\mathcal{N})}(a)$ *with* $* \in \{\min, \max, \sum, \varnothing, \varnothing_{hyb}, ML\}$.

2. $\widetilde{SV}^{*}_{\mathsf{T}(\mathcal{N})}(a) = SV^{*}_{\mathsf{T}(\mathcal{N})}(a) - \frac{PD(\{a\})}{n}$ *with* $* \in \{\min, \max, \varnothing, \varnothing_{hyb}, ML\}$.

3. $\widetilde{SV}^{\sum}_{\mathsf{T}(\mathcal{N})}(a) = SV^{\sum}_{\mathsf{T}(\mathcal{N})}(a) - |\,\mathsf{T}(\mathcal{N})\,| \cdot \frac{PD(\{a\})}{n}$.

(In case of $FP^{ML}_{\mathsf{T}(\mathcal{N})}$, $SV^{ML}_{\mathsf{T}(\mathcal{N})}$ and $\widetilde{SV}^{ML}_{\mathsf{T}(\mathcal{N})}$, we assume the most likely embedded tree to be fixed, if the argmax is not unique. Otherwise, i.e. if different most likely embedded trees are considered when calculating the Fair Proportion Index or the different versions of the Shapley Value, we cannot assess their relationship).

Proof. The first two properties follow directly from the corresponding properties of the *Fair Proportion Index*, the *Shapley Value* and the *modified Shapley Value* on phylogenetic trees (cf. Summary on page 13). To be precise, we have

- $FP_{\mathcal{T}}(a) = SV_{\mathcal{T}}(a)$ and
- $\widetilde{SV}_{\mathcal{T}}(a) = SV_{\mathcal{T}}(a) - \frac{PD(\{a\})}{n}$

for all embedded trees $\mathcal{T} \in \mathsf{T}(\mathcal{N})$. Thus, $FP^{*}_{\mathsf{T}(\mathcal{N})}(a) = SV^{*}_{\mathsf{T}(\mathcal{N})}(a)$ with $* \in \{\min, \max, \sum, \varnothing, \varnothing_{hyb}, ML\}$, because all functions $*$ have the same input for FP and SV.
With the exception of $* = \sum$, the same reasoning yields

$$\widetilde{SV}^{*}_{\mathsf{T}(\mathcal{N})}(a) = SV^{*}_{\mathsf{T}(\mathcal{N})}(a) - \frac{PD(\{a\})}{n},$$

because the corresponding relationship between $\widetilde{SV}(a)$ and $SV(a)$ on all embedded trees $\mathcal{T} \in \mathsf{T}(\mathcal{N})$ is preserved by $* \in \{\min, \max, \varnothing, \varnothing_{hyb}, ML\}$.
In case of $\widetilde{SV}^{\sum}_{\mathsf{T}(\mathcal{N})}(a)$ and $SV^{\sum}_{\mathsf{T}(\mathcal{N})}(a)$, however, the *original Shapley Values* $SV(a)$ and *modified Shapley Values* $\widetilde{SV}(a)$ are added for all embedded trees $\mathcal{T} \in \mathsf{T}(\mathcal{N})$, respectively. Thus,

$$\begin{aligned}
\widetilde{SV}^{\sum}_{\mathsf{T}(\mathcal{N})}(a) &= \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} \widetilde{SV}_{\mathcal{T}}(a) \\
&= \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} \Big(SV_{\mathcal{T}}(a) - \frac{PD(\{a\})}{n}\Big) \\
&= -|\,\mathsf{T}(\mathcal{N})\,| \cdot \frac{PD(\{a\})}{n} + \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} SV_{\mathcal{T}}(a) \\
&= -|\,\mathsf{T}(\mathcal{N})\,| \cdot \frac{PD(\{a\})}{n} + SV^{\sum}_{\mathsf{T}(\mathcal{N})}(a) \\
&= SV^{\sum}_{\mathsf{T}(\mathcal{N})}(a) - |\,\mathsf{T}(\mathcal{N})\,| \cdot \frac{PD(\{a\})}{n}.
\end{aligned}$$

□

Similar to the results for phylogenetic trees, Proposition 5 suggests that the infeasible calculation of the *embedded original Shapley Value* can be replaced by the simpler calculation of the *embedded Fair Proportion Index*. Accordingly, the *embedded modified*

Shapley Value can be derived from the *embedded Fair Proportion Index*. The *embedded unrooted rooted Shapley Value*, on the other hand, cannot be derived from any of the other indices. It can be computed according to its definition or via so-called *splits* induced by the edges of the tree (cf. page 110), which is less complex.

However, both the *embedded Fair Proportion Index* and all versions of the *embedded Shapley Value* are based on the set $\mathsf{T}(\mathcal{N})$ of phylogenetic X-trees displayed by $\mathcal{N}$. As the determination of $\mathsf{T}(\mathcal{N})$ is an NP-hard problem, so is the computation of any embedded biodiversity index.

5.2.2. LSA associated Shapley Value

Similar to Chapter 5.1.2 we will now introduce the *LSA associated Shapley Value* for a phylogenetic network $\mathcal{N}$ on X.

Definition 38 (LSA associated Shapley Value)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on some taxon set X. Let $a \in X$ be a taxon in X. Then we define

$$SV^{LSA}(a) := SV_{\mathcal{T}_{LSA}(\mathcal{N})}(a), \tag{5.19}$$

$$\widetilde{SV}^{LSA}(a) := \widetilde{SV}_{\mathcal{T}_{LSA}(\mathcal{N})}(a) \text{ and} \tag{5.20}$$

$$\widehat{SV}^{LSA}(a) := \widehat{SV}_{\mathcal{T}_{LSA}(\mathcal{N})}(a), \tag{5.21}$$

where $SV_{\mathcal{T}_{LSA}(\mathcal{N})}(a)$ is the *original Shapley Value* of a, $\widetilde{SV}_{\mathcal{T}_{LSA}(\mathcal{N})}(a)$ is the *modified Shapley Value* of a and $\widehat{SV}_{\mathcal{T}_{LSA}(\mathcal{N})}(a)$ is the *unrooted rooted Shapley Value* of a in the *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N})$ associated with $\mathcal{N}$. If hybridization probabilities are given for $\mathcal{N}$, we also consider

$$SV^{LSA_{hyb}}(a) := SV_{\mathcal{T}^{hyb}_{LSA}(\mathcal{N})}(a), \tag{5.22}$$

$$SV^{LSA_{ML}}(a) := SV_{\mathcal{T}^{ML}_{LSA}(\mathcal{N})}(a), \tag{5.23}$$

$$\widetilde{SV}^{LSA_{hyb}}(a) := \widetilde{SV}_{\mathcal{T}^{hyb}_{LSA}(\mathcal{N})}(a), \tag{5.24}$$

$$\widetilde{SV}^{LSA_{ML}}(a) := \widetilde{SV}_{\mathcal{T}^{ML}_{LSA}(\mathcal{N})}(a), \tag{5.25}$$

$$\widehat{SV}^{LSA_{hyb}}(a) := \widehat{SV}_{\mathcal{T}^{hyb}_{LSA}(\mathcal{N})}(a), \tag{5.26}$$

$$\widehat{SV}^{LSA_{ML}}(a) := \widehat{SV}_{\mathcal{T}^{ML}_{LSA}(\mathcal{N})}(a), \tag{5.27}$$

where the *original Shapley Value*, the *modified Shapley Value* and the *unrooted rooted Shapley Value* of a are calculated for the *hybrid LSA tree* and the *Maximum Likelihood LSA tree*, respectively. We will call all SV^* with $* \in \{LSA, LSA_{hyb}, LSA_{ML}\}$ the *LSA associated original Shapley Value*, $\widetilde{SV}^*$ the *LSA associated modified Shapley Value* and

$\widehat{SV}^*$ the *LSA associated unrooted rooted Shapley Value* and use the superscript to indicate which version of the *LSA tree* was used.

Example 19. Consider the rooted phylogenetic network $\mathcal{N}_8$ on $X = \{A, B, C, D\}$ and its *hybrid LSA tree* depicted in Figure 17. For $\mathcal{N}_8$ all versions of the *LSA tree* coincide, i.e.

$$\mathcal{T}_{LSA}(\mathcal{N}_8) = \mathcal{T}_{LSA}^{hyb}(\mathcal{N}_8) = \mathcal{T}_{LSA}^{ML}(\mathcal{N}).$$

Let $* \in \{LSA, LSA_{hyb}, LSA_{ML}\}$. Then we retrieve the following *LSA associated Shapley Values*:

- *LSA associated original Shapley Value*:

$$\begin{aligned} SV^*(A) &= 3 \\ \widehat{SV}^*(A) &= 3 \\ SV^*(B) &= 3 \\ SV^*(C) &= 3 \\ SV^*(D) &= 3. \end{aligned}$$

- *LSA associated modified Shapley Value:*

$$\begin{aligned} \widetilde{SV}^*(A) &= \frac{9}{4} \\ \widetilde{SV}^*(B) &= \frac{9}{4} \\ \widetilde{SV}^*(C) &= \frac{9}{4} \\ \widetilde{SV}^*(D) &= \frac{9}{4}. \end{aligned}$$

- *LSA associated unrooted rooted Shapley Value:*

$$\begin{aligned} \widehat{SV}^*(A) &= 3 \\ \widehat{SV}^*(B) &= 3 \\ \widehat{SV}^*(C) &= 3 \\ \widehat{SV}^*(D) &= 3. \end{aligned}$$

By chance the *LSA associated original Shapley Value* and the *LSA associated unrooted rooted Shapley Value* coincide in this example. In general, however, the two indices differ.

As the *LSA associated Fair Proportion Index* and the different versions of the *LSA associated Shapley Value* are calculated by considering the *LSA tree* associated with a network, we have the following corollary:

Corollary 1. *Let $\mathcal{N}$ be a phylogenetic network on X with $|X| = n$ and let $a \in X$ be a taxon of $\mathcal{N}$. Then we have*

1. $FP^*(a) = SV^*(a)$;
2. $\widetilde{SV}^*(a) = SV^*(a) - \frac{PD(\{a\})}{n}$,

with $* \in \{LSA, LSA_{hyb}, LSA_{ML}\}$.

Proof. Both claims follow directly from the corresponding properties of the *Fair Proportion Index*, the *original Shapley Value* and the *modified Shapley Value* on phylogenetic trees. □

5.2.3. Generalized Shapley Value

After reducing a rooted phylogenetic network $\mathcal{N}$ to its treelike content in order to calculate the different versions of the *Shapley Value* for its taxa, we now directly calculate the indices according to their definition (cf. Definitions 6, 7) by using the different measures of *generalized phylogenetic diversity* introduced in Chapter 4.

However, as we have only defined *phylogenetic diversity* for rooted phylogenetic networks, we will not consider the *unrooted rooted Shapley Value* in the following, but only the *original Shapley Value* and the *modified Shapley Value*.

Definition 39 (Generalized Shapley Value)**.** Let $\mathcal{N}$ be a rooted phylogenetic network on some taxon set X and let $\mathsf{T}(\mathcal{N})$ be the (multi)set of all rooted phylogenetic X-trees displayed by $\mathcal{N}$. Let $a \in X$ be a taxon in X and let $\mathcal{PD}(S)$ denote any generalized measure of *phylogenetic diversity* of a subset $S \subseteq X$ of taxa in $\mathcal{N}$, i.e. $\mathcal{PD}(S) \in \{PND(S), PND^{hyb}(S), PND^{ML}(S), PD^{\min}_{\mathsf{T}(\mathcal{N})}(S), PD^{\max}_{\mathsf{T}(\mathcal{N})}(S), PD^{\sum}_{\mathsf{T}(\mathcal{N})}(S), PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}(S), PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(S), PD^{ML}_{\mathsf{T}(\mathcal{N})}(S), PD^{LSA}(S), PD^{LSA_{hyb}}(S), PD^{LSA_{ML}}(S)\}$.
Then we define the *generalized original Shapley Value* of a as

$$SV_{\mathcal{PD}}(a) = \frac{1}{n!} \sum_{\substack{S \subseteq X \\ a \in S}} \Big((|S| - 1)!(n - |S|)!(\mathcal{PD}(S) - \mathcal{PD}(S \setminus \{a\})) \Big), \tag{5.28}$$

where $n = |X|$ and S denotes a subset of species containing taxon a and the sum runs over all such subsets possible.

Analogously, we define the *generalized modified Shapley Value* of a as

$$\widetilde{SV}_{\mathcal{PD}}(a) = \frac{1}{n!} \sum_{\substack{S: a \in S \\ |S| \geq 2}} \Big((|S|-1)!(n-|S|)!(\mathcal{PD}(S) - \mathcal{PD}(S \setminus \{a\})) \Big), \tag{5.29}$$

where the sum runs over all coalitions S containing taxon a and at least one other taxon.

Remark. Obviously, we have

$$\widetilde{SV}_{\mathcal{PD}}(a) = SV_{\mathcal{PD}}(a) - \frac{\mathcal{PD}(a)}{n},$$

for all $a \in X$ (cf. Proposition 1).

Example 20. Consider the rooted phylogenetic network $\mathcal{N}_8$ on $X = \{A, B, C, D\}$ depicted in Figure 23. We now calculate the *generalized original Shapley Value* of taxon $A \in X$ and choose the *phylogenetic net diversity* as the measure of *phylogenetic diversity* (cf. Table 7 for the values of PND). We have to consider the following subsets $S \subseteq X$: $\{A\}, \{A, B\}, \{A, C\}, \{A, D\}, \{A, B, C\}, \{A, B, D\}, \{A, C, D\}$ and $\{A, B, C, D\}$. Thus,

$$\begin{aligned}
SV_{PND}(A) &= \frac{1}{4!} \sum_{S: A \in S} \Big((|S|-1)!(|X|-|S|)!(PND(S) - PND(S \setminus \{A\})) \Big) \\
&= \frac{1}{4!} \Big[(1-1)!(4-1)!(3-0) \\
&\quad + (2-1)!(4-2)!\big((5-3)+(5-3)+(6-3)\big) \\
&\quad + (3-1)!(4-3)!\big((6-4)+(8-5)+(7-4)\big) \\
&\quad + (4-1)!(4-4)!(9-6) \Big] \\
&= \frac{1}{24} \Big[1 \cdot 6 \cdot 3 + 1 \cdot 2 \cdot (2+2+3) + 2 \cdot 1 \cdot (2+3+3) + 6 \cdot 1 \cdot 3 \Big] \\
&= \frac{1}{24} \cdot 66 \\
&= \frac{11}{4}.
\end{aligned}$$

For the *generalized modified Shapley Value* we have

$$\begin{aligned}
\widetilde{SV}_{PND}(A) &= \frac{1}{4!}\sum_{\substack{S:a\in S\\ |S|\geq 2}}\Big((|S|-1)!(n-|S|)!(PND(S)-PND(S\setminus\{a\}))\Big)\\
&= \frac{1}{4!}\Big[(2-1)!(4-2)!\big((5-3)+(5-3)+(6-3)\big)\\
&\quad +(3-1)!(4-3)!\big((6-4)+(8-5)+(7-4)\big)\\
&\quad +(4-1)!(4-4)!(9-6)\Big]\\
&= \frac{1}{24}\Big[1\cdot 2\cdot(2+2+3)+2\cdot 1\cdot(2+3+3)+6\cdot 1\cdot 3\Big]\\
&= \frac{1}{24}\cdot 48\\
&= 2.
\end{aligned}$$

Analogously, the *generalized original Shapley Value* can be calculated for all other taxa in X and all other generalized measures of *phylogenetic diversity*. Table 7 summarizes the results.

Note that for all subsets $S \subseteq X$ in $\mathcal{N}_8$ we have:

- $PND(S) = PD^{\min}_{\mathsf{T}(\mathcal{N}_8)}(S)$;
- $PND^{hyb}(S) = PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_8)}(S)$;
- $PND^{ML}(S) = PD^{ML}_{\mathsf{T}(\mathcal{N}_8)}(S)$.

Thus, the corresponding *generalized (modified) Shapley Values* coincide for these measures, respectively. However, in general, they differ. To see this consider, for example, the rooted phylogenetic network $\mathcal{N}_2'$ depicted Figure 16 and some of its *generalized Shapley Values* listed in Table 8.

Table 7: *Generalized phylogenetic diversity* and *generalized (modified) Shapley Values* for the rooted phylogenetic network $\mathcal{N}_8$

S	PND	PND^{hyb}	PND^{ML}	$PD^{\min}_{\mathsf{T}(\mathcal{N}_8)}$	$PD^{\max}_{\mathsf{T}(\mathcal{N}_8)}$	$PD^{\sum}_{\mathsf{T}(\mathcal{N}_8)}$	$PD^{\varnothing}_{\mathsf{T}(\mathcal{N}_8)}$	$PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_8)}$	$PD^{ML}_{\mathsf{T}(\mathcal{N}_8)}$	PD^{LSA}	$PD^{LSA_{hyb}}$	$PD^{LSA_{ML}}$
$\emptyset$	0	0	0	0	0	0	0	0	0	0	0	0
$\{A\}$	3	3	3	3	3	12	3	3	3	3	3	3
$\{B\}$	3	3	3	3	3	12	3	3	3	3	3	3
$\{C\}$	3	3	3	3	3	12	3	3	3	3	3	3
$\{D\}$	3	3	3	3	3	12	3	3	3	3	3	3
$\{A,B\}$	5	$\frac{17}{3}$	6	5	6	22	$\frac{11}{2}$	$\frac{17}{3}$	6	6	6	6
$\{A,C\}$	5	$\frac{23}{4}$	6	5	6	23	$\frac{23}{4}$	$\frac{23}{4}$	6	6	6	6
$\{A,D\}$	6	6	6	6	6	24	6	6	6	6	6	6
$\{B,C\}$	4	$\frac{13}{3}$	4	4	6	19	$\frac{19}{4}$	$\frac{13}{3}$	4	6	6	6
$\{B,D\}$	5	$\frac{16}{3}$	5	5	6	22	$\frac{11}{2}$	$\frac{16}{3}$	5	6	6	6
$\{C,D\}$	4	5	5	4	6	19	$\frac{19}{4}$	5	5	6	6	6
$\{A,B,C\}$	6	7	7	6	8	29	$\frac{29}{4}$	7	7	9	9	9
$\{A,B,D\}$	8	8	8	8	8	32	8	8	8	9	9	9
$\{A,C,D\}$	7	$\frac{31}{4}$	8	7	8	30	$\frac{15}{2}$	$\frac{31}{4}$	8	9	9	9
$\{B,C,D\}$	6	$\frac{19}{3}$	6	6	7	26	$\frac{13}{2}$	$\frac{19}{3}$	6	9	9	9
$\{A,B,C,D\}$	9	9	9	9	9	36	9	9	9	12	12	12
$SV_{\mathcal{PD}}(A)$	$\frac{11}{4}$	$\frac{67}{24}$	3	$\frac{11}{4}$	$\frac{5}{2}$	$\frac{65}{6}$	$\frac{65}{24}$	$\frac{67}{24}$	3	3	3	3
$SV_{\mathcal{PD}}(B)$	$\frac{25}{12}$	$\frac{71}{36}$	$\frac{11}{6}$	$\frac{25}{12}$	$\frac{13}{6}$	$\frac{17}{2}$	$\frac{17}{8}$	$\frac{71}{36}$	$\frac{11}{6}$	3	3	3
$SV_{\mathcal{PD}}(C)$	$\frac{19}{12}$	$\frac{133}{72}$	$\frac{11}{6}$	$\frac{19}{12}$	$\frac{13}{6}$	$\frac{15}{2}$	$\frac{15}{8}$	$\frac{133}{72}$	$\frac{11}{6}$	3	3	3
$SV_{\mathcal{PD}}(D)$	$\frac{31}{12}$	$\frac{43}{18}$	$\frac{7}{3}$	$\frac{31}{12}$	$\frac{13}{6}$	$\frac{55}{6}$	$\frac{55}{24}$	$\frac{43}{18}$	$\frac{7}{3}$	3	3	3
$\widetilde{SV}_{\mathcal{PD}}(A)$	2	$\frac{49}{24}$	$\frac{9}{4}$	2	$\frac{7}{4}$	$\frac{47}{6}$	$\frac{47}{24}$	$\frac{49}{24}$	$\frac{9}{4}$	$\frac{9}{4}$	$\frac{9}{4}$	$\frac{9}{4}$
$\widetilde{SV}_{\mathcal{PD}}(B)$	$\frac{4}{3}$	$\frac{11}{9}$	$\frac{13}{12}$	$\frac{4}{3}$	$\frac{17}{12}$	$\frac{11}{2}$	$\frac{11}{8}$	$\frac{11}{9}$	$\frac{13}{12}$	$\frac{9}{4}$	$\frac{9}{4}$	$\frac{9}{4}$
$\widetilde{SV}_{\mathcal{PD}}(C)$	$\frac{5}{6}$	$\frac{79}{72}$	$\frac{13}{12}$	$\frac{5}{6}$	$\frac{17}{12}$	$\frac{9}{2}$	$\frac{9}{8}$	$\frac{79}{72}$	$\frac{13}{12}$	$\frac{9}{4}$	$\frac{9}{4}$	$\frac{9}{4}$
$\widetilde{SV}_{\mathcal{PD}}(D)$	$\frac{11}{6}$	$\frac{59}{36}$	$\frac{19}{12}$	$\frac{11}{6}$	$\frac{17}{12}$	$\frac{37}{6}$	$\frac{37}{24}$	$\frac{59}{36}$	$\frac{19}{12}$	$\frac{9}{4}$	$\frac{9}{4}$	$\frac{9}{4}$

Table 8: *Generalized phylogenetic diversity* and *generalized Shapley Values* for the rooted phylogenetic network $\mathcal{N}_2'$ (extract)

S	$PD^{\min}_{\mathsf{T}(\mathcal{N}_2')}$	$PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N}_2')}$	$PD^{ML}_{\mathsf{T}(\mathcal{N}_2')}$	PND	PND^{hyb}	PND^{ML}
$\emptyset$	0	0	0	0	0	0
$\{A\}$	3	3	3	3	3	3
$\{B\}$	3	3	3	3	3	3
$\{C\}$	3	3	3	3	3	3
$\{D\}$	3	3	3	3	3	3
$\{A,B\}$	4	$\frac{9}{2}$	4	4	$\frac{17}{4}$	4
$\{A,C\}$	5	$\frac{16}{3}$	5	5	$\frac{17}{3}$	6
$\{A,D\}$	6	6	6	6	6	6
$\{B,C\}$	4	$\frac{31}{6}$	5	4	$\frac{67}{12}$	6
$\{B,D\}$	6	6	6	6	6	6
$\{C,D\}$	4	$\frac{16}{3}$	6	4	$\frac{14}{3}$	4
$\{A,B,C\}$	6	$\frac{20}{3}$	6	6	$\frac{41}{6}$	7
$\{A,B,D\}$	7	$\frac{15}{2}$	7	7	$\frac{29}{4}$	7
$\{A,C,D\}$	7	$\frac{23}{3}$	8	7	$\frac{22}{3}$	7
$\{B,C,D\}$	7	$\frac{15}{2}$	8	7	$\frac{29}{4}$	7
$\{A,B,C,D\}$	9	9	9	8	$\frac{17}{2}$	8
$SV_{\mathcal{PD}}(A)$	$\frac{9}{4}$	$\frac{77}{36}$	$\frac{11}{6}$	2	$\frac{149}{72}$	2
$SV_{\mathcal{PD}}(B)$	$\frac{25}{12}$	$\frac{37}{18}$	$\frac{11}{6}$	$\frac{11}{6}$	$\frac{73}{36}$	2
$SV_{\mathcal{PD}}(C)$	$\frac{23}{12}$	$\frac{77}{36}$	$\frac{7}{3}$	$\frac{5}{3}$	$\frac{149}{72}$	2
$SV_{\mathcal{PD}}(D)$	$\frac{11}{4}$	$\frac{8}{3}$	3	$\frac{5}{2}$	$\frac{7}{3}$	2

Relationship between the generalized Shapley Value and the embedded Shapley Value

We now shortly compare the *generalized original Shapley Value* and the *embedded original Shapley Value* of a phylogenetic network $\mathcal{N}$ on some taxon set X.

The first observation to make is that, in general,

- $SV_{PD^{\min}_{\mathsf{T}(\mathcal{N})}}(a) \neq SV^{\min}_{\mathsf{T}(\mathcal{N})}(a)$ and
- $SV_{PD^{\max}_{\mathsf{T}(\mathcal{N})}}(a) \neq SV^{\max}_{\mathsf{T}(\mathcal{N})}(a)$

for $a \in X$. Consider for example the rooted phylogenetic network $\mathcal{N}_8$ depicted in Figure 23. Here we have $SV^{\min}_{\mathsf{T}(\mathcal{N}_8)}(A) = \frac{7}{3}$ (cf. Table 4), but $SV_{PD^{\min}_{\mathsf{T}(\mathcal{N}_8)}}(A) = \frac{11}{4}$ (cf. Table 7). Similarly, we have $SV^{\max}_{\mathsf{T}(\mathcal{N}_8)}(A) = 3$ (cf. Table 4), but $SV_{PD^{\max}_{\mathsf{T}(\mathcal{N}_8)}}(A) = \frac{5}{2}$ (cf. Table 7). Analogous results hold for the *generalized modified Shapley Value* and the *embedded modified Shapley Value*, i.e. in general

- $\widetilde{SV}_{PD^{\min}_{\mathsf{T}(\mathcal{N})}}(a) \neq \widetilde{SV}^{\min}_{\mathsf{T}(\mathcal{N})}(a)$ and
- $\widetilde{SV}_{PD^{\max}_{\mathsf{T}(\mathcal{N})}}(a) \neq \widetilde{SV}^{\max}_{\mathsf{T}(\mathcal{N})}(a)$.

The second observation to make is

- $SV_{PD^{ML}_{\mathsf{T}(\mathcal{N})}}(a) = SV^{ML}_{\mathsf{T}(\mathcal{N})}(a)$ and
- $\widetilde{SV}_{PD^{ML}_{\mathsf{T}(\mathcal{N})}}(a) = \widetilde{SV}^{ML}_{\mathsf{T}(\mathcal{N})}(a)$

if the most likely tree $\mathcal{T}^* \in \mathsf{T}(\mathcal{N}) = \operatorname*{argmax}_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} \mathbb{P}(\mathcal{T})$ is fixed, because:

$$\begin{aligned} SV^{ML}_{\mathsf{T}(\mathcal{N})}(a) &= SV_{\mathcal{T}^*} \quad \text{with } \mathcal{T}^* = \operatorname*{argmax}_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} \mathbb{P}(\mathcal{T}) \\ &= \frac{1}{n!} \sum_{\substack{S\subseteq X \\ a\in S}} \Big((|S|-1)!(n-|S|)!(PD_{\mathcal{T}^*}(S) - PD_{\mathcal{T}^*}(S\setminus\{a\})) \Big) \\ &= SV_{PD^{ML}_{\mathsf{T}(\mathcal{N})}}(a). \end{aligned}$$

Analogously, the equality follows for the *modified Shapley Values*.

Lastly, we have the following relationship:

Proposition 6. *Let $\mathcal{N}$ be a rooted phylogenetic network on some taxon set X with $|X| = n$ and let $\mathsf{T}(\mathcal{N})$ be the (multi)set of all rooted phylogenetic X-trees displayed by $\mathcal{N}$. Let $a \in X$ be a taxon in X. Then*

1. $SV_{PD^{\sum}_{\mathsf{T}(\mathcal{N})}}(a) = SV^{\sum}_{\mathsf{T}(\mathcal{N})}(a)$,
2. $SV_{PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}}(a) = SV^{\varnothing}_{\mathsf{T}(\mathcal{N})}(a)$,
3. $SV_{PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}}(a) = SV^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}(a)$.

Analogously, the corresponding embedded and generalized modified Shapley Values coincide.

Proof. We only show 1., but 2. and 3. follow analogously.
Recall that $PD^{\sum}_{\mathsf{T}(\mathcal{N})}(S) = \sum\limits_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} PD_{\mathcal{T}}(S)$. Thus,

$$\begin{aligned}
SV_{PD^{\sum}_{\mathsf{T}(\mathcal{N})}}(a) &= \frac{1}{n!}\sum_{\substack{S\subseteq X\\ a\in S}}\Big((|S|-1)!(n-|S|)!(PD^{\sum}_{\mathsf{T}(\mathcal{N})}(S) - PD^{\sum}_{\mathsf{T}(\mathcal{N})}(S\setminus\{a\}))\Big)\\
&= \frac{1}{n!}\sum_{\substack{S\subseteq X\\ a\in S}}\Big((|S|-1)!(n-|S|)!\big(\sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} PD_{\mathcal{T}}(S) - \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} PD_{\mathcal{T}}(S\setminus\{a\})\big)\Big)\\
&= \frac{1}{n!}\sum_{\substack{S\subseteq X\\ a\in S}}\Big((|S|-1)!(n-|S|)!\big(\sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} (PD_{\mathcal{T}}(S) - PD_{\mathcal{T}}(S\setminus\{a\}))\big)\Big).
\end{aligned}$$

On the other hand we have

$$\begin{aligned}
SV^{\sum}_{\mathsf{T}(\mathcal{N})}(a) &= \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} SV_{\mathcal{T}}(a)\\
&= \sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})}\Big(\frac{1}{n!}\sum_{\substack{S\subseteq X\\ a\in S}}\big((|S|-1)!(n-|S|)!(PD_{\mathcal{T}}(S) - PD_{\mathcal{T}}(S\setminus\{a\}))\big)\Big)\\
&= \frac{1}{n!}\sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})}\sum_{\substack{S\subseteq X\\ a\in S}}\Big((|S|-1)!(n-|S|)!(PD_{\mathcal{T}}(S) - PD_{\mathcal{T}}(S\setminus\{a\}))\Big)\\
&= \frac{1}{n!}\sum_{\substack{S\subseteq X\\ a\in S}}\sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})}\Big((|S|-1)!(n-|S|)!(PD_{\mathcal{T}}(S) - PD_{\mathcal{T}}(S\setminus\{a\}))\Big)\\
&= \frac{1}{n!}\sum_{\substack{S\subseteq X\\ a\in S}}\Big((|S|-1)!(n-|S|)!\big(\sum_{\mathcal{T}\in\mathsf{T}(\mathcal{N})} (PD_{\mathcal{T}}(S) - PD_{\mathcal{T}}(S\setminus\{a\}))\big)\Big).
\end{aligned}$$

Thus,

$$SV_{PD^{\sum}_{\mathsf{T}(\mathcal{N})}}(a) = SV^{\sum}_{\mathsf{T}(\mathcal{N})}(a).$$

□

Remark. Above equalities together with Proposition 5 suggest to derive $SV_{PD^{\sum}_{\mathsf{T}(\mathcal{N})}}$, $SV_{PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}}$, $SV_{PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}}$ and $SV_{PD^{ML}_{\mathsf{T}(\mathcal{N})}}$ (and analogously $\widetilde{SV}_{PD^{\sum}_{\mathsf{T}(\mathcal{N})}}$, etc.) from the corresponding versions of the *embedded Fair Proportion Index*. For SV_{PND}, $SV_{PD^{\min}_{\mathsf{T}(\mathcal{N})}}$ and $SV_{PD^{\min}_{\mathsf{T}(\mathcal{N})}}$ (and analogously $\widetilde{SV}_{PND}$, etc.), however, this is not possible. Thus, these versions of the *generalized original Shapley Value* have to be calculated according to their definition.

Relationship between the generalized Shapley Value and the LSA associated Shapley Value

Both the calculation of the *LSA associated Shapley Value* SV^{LSA} and the *generalized Shapley Value* $SV_{PD^{LSA}}$, that uses the *LSA associated phylogenetic diversity*, are based upon the same *LSA tree* associated with a phylogenetic network $\mathcal{N}$ on X. Analogously, this holds for the *hybrid LSA tree* and the *Maximum Likelihood LSA Tree*. Thus, we have the following relationship for a taxon $a \in X$.

- $\mathcal{SV}^{LSA}(a) = \mathcal{SV}_{PD^{LSA}}(a)$ with $\mathcal{SV} \in \{SV, \widetilde{SV}\}$;
- $\mathcal{SV}^{LSA_{hyb}}(a) = \mathcal{SV}_{PD^{LSA_{hyb}}}(a)$ with $\mathcal{SV} \in \{SV, \widetilde{SV}\}$;
- $\mathcal{SV}^{LSA_{ML}}(a) = \mathcal{SV}_{PD^{LSA_{ML}}}(a)$ with $\mathcal{SV} \in \{SV, \widetilde{SV}\}$.

Moreover, $\mathcal{SV}^{LSA}$, $\mathcal{SV}^{LSA_{hyb}}$ and $\mathcal{SV}^{LSA_{ML}}$ with $\mathcal{SV} \in \{SV, \widetilde{SV}\}$ can be derived from the corresponding versions of the *LSA associated Fair Proportion Index* due to Corollary 1.

5.3. Conclusion

In order to generalize the *Fair Proportion Index* and the different versions of the *Shapley Value* from phylogenetic trees to phylogenetic networks, we have considered different approaches.

For the *Fair Proportion Index*, we have on the one hand considered the treelike content of a phylogenetic network via the (multi)set $\mathsf{T}(\mathcal{N})$ of embedded trees and the *LSA tree* associated with the network, and on the other hand, we have developed a measure (the *Net Fair Proportion Index*) very similar to the *Fair Proportion Index* for phylogenetic trees, but directly based on the phylogenetic network. All of these

approaches have their specific advantages and drawbacks. While the *Fair Proportion Index* is easy to calculate for a single phylogenetic tree, the task of calculating any version of the *embedded Fair Proportion Index* based on the (multi)set $\mathsf{T}(\mathcal{N})$ may be difficult to accomplish in practice, because the determination of the (multi)set $\mathsf{T}(\mathcal{N})$ is an NP-hard problem (cf. Linz et al. [25]). Still, in particular the $FP^{\varnothing}_{\mathsf{T}(\mathcal{N})}$ and the $FP^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}$ seem to be the most sensible versions of the generalized *Fair Proportion Index* to be used, because they take into account the most information provided by the network.

The *LSA associated Fair Proportion Index* on the other hand, reduces the phylogenetic network to its most basic treelike content and thus, discards a lot of the evolutionary information present. It may therefore be questionable if the *LSA associated Fair Proportion Index* is an appropriate measure to be used in taxon prioritization. On the other hand, this measure seems to be more easy to calculate in practice than the *embedded Fair Proportion Index* (cf. Chapter 6.3), even though it is possible to construct examples where the construction of the *LSA tree* prior to the calculation of the *LSA associated Fair Proportion Index* faces problems, because of an exponential number of paths between a reticulation node and its *lowest stable ancestor* (cf. Remark on page 28).

The problem of an exponential number of paths between a reticulation node and its *lowest stable ancestor* may also affect the calculation of the *Net Fair Proportion Index* NFP, because the calculation of this measures requires the enumeration of all paths between the root and each of the leaves (cf. Remarks on page 77). Moreover, the *Net Fair Proportion Index* is a relatively arbitrary measure and lacks a link to biological conservation. Its idea is to divide all branch lengths equally among the taxa of the network. Thus, it directly resembles the *Fair Proportion Index* for phylogenetic trees, which has also been criticized of lacking a biological motivation, but in the meantime has been justified by its equality with the *original Shapley Value*. It may therefore be worth investigating the relationship of the *Net Fair Proportion Index* with any version of the *Shapley Value* for phylogenetic networks or any other biodiversity index for phylogenetic networks in order to assess the suitability of the NFP as a taxon prioritization criterion.

Before analyzing the relationship between the *Net Fair Proportion Index* and the *Shapley Value* for phylogenetic networks, it is, however, necessary to assess the concept of the *Shapley Value* for phylogenetic networks itself. In order to generalize the concept of the different versions of the *Shapley Value* from phylogenetic trees to networks, we have, again, considered different approaches.

On the one hand, we have reduced the phylogenetic network to its treelike content via

the (multi)set $\mathsf{T}(\mathcal{N})$ of embedded trees and the *LSA tree* associated with the network and have defined the different versions of the *Shapley Value* based on these trees. Completely in accordance with the *embedded Fair Proportion Index* and the *LSA associated Fair Proportion Index*, both the *embedded Shapley Value* and the different versions of he *LSA associated Shapley Value* have their specific advantages and drawbacks. While in particular $\mathcal{SV}^{\varnothing}_{\mathsf{T}(\mathcal{N})}$ and $\mathcal{SV}^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}$ with $\mathcal{SV} \in \{SV, \widetilde{SV}, \widehat{SV}\}$ seem sensible to use, the different versions of the *LSA associated Shapley Value* are questionable. In all cases, however, we suggest to derive the different versions of the *embedded (original/modified) Shapley Value* and the *LSA associated (original/modified) Shapley Value* from the corresponding versions of the *embedded Fair Proportion Index* and the *LSA associated Fair Proportion Index*.

However, as the different versions of the *Shapley Value* are based on the *phylogenetic diversity* of subsets of taxa and thus do not directly depend on the structure of a phylogenetic tree or network, we have introduced the *generalized (original/modified) Shapley Value* as an alternative approach. This index uses any measure of generalized *phylogenetic diversity* to calculate the *(original/modified) Shapley Value* directly from its definition. However, this has the drawback that the chosen measure of generalized *phylogenetic diversity* has to be calculated for all $2^{|X|}$ subsets $S \subseteq X$. Thus, not only may the calculation of the chosen measure of generalized *phylogenetic diversity* in itself be a limiting factor in practice, but also the high number of subsets to be considered. However, some of versions of the *(original/modified) Shapley Value* can be derived from the *embedded Fair Proportion Index*, which eases its calculation to some extent, even tough we remark again, that the calculation of any embedded biodiversity index is an NP-hard problem due to the required determination of the (multi)set $\mathsf{T}(\mathcal{N})$.

All in all, we have considered a variety of approaches towards the generalization of the *Fair Proportion Index* and the different versions of the *Shapley Value* from phylogenetic trees to networks. All of these approaches have their specific advantages and drawbacks. In particular, computational feasibility may be a problem in practice for some of the indices. On the other hand, biological plausibility and suitability for taxon prioritization remain to be assessed for all indices. It may therefore be worth analyzing the relationship between the individual indices and evaluating their biological meaning. Once promising indices have been identified, it may then be necessary to develop approximations for these indices, in order to use them in practical applications.

In the following, we will, however, introduce the software tool `net_diversity.pl` that allows for the calculation of some of the generalized biodiversity indices introduced in this chapter.

6. Software

6.1. Extended Newick Format

In order to represent phylogenetic networks in a standardized and computer-readable format, the *extended Newick format* was introduced as a generalization of the *Newick tree format*. Before we go into more detail about the former, we shortly recapitulate the latter.

The Newick tree format (cf. Felsenstein et al. [12]) describes a phylogenetic tree as a string using nested parentheses and commas, where closely related species are grouped closely together.

Leaves are represented by their names and each pair of matched parentheses represents an internal node, which may additionally have a name assigned to it.

Branch lengths can optionally be included in the representation by putting a colon followed by a real number after a node.

Every tree ends with a semicolon.

When the Newick tree format is used to represent an unrooted phylogenetic tree, an arbitrary inner node is chosen as its root, because the Newick tree format is fashioned to represent rooted trees and requires a root node represented by the outmost pair of parentheses. In case of binary phylogenetic trees it is, however, still possible to distinguish between rooted and unrooted trees. Whereas a rooted binary tree has two entries at each parentheses level, an unrooted binary tree has two entries at each parentheses level, except for the uppermost level, where it has three entries.

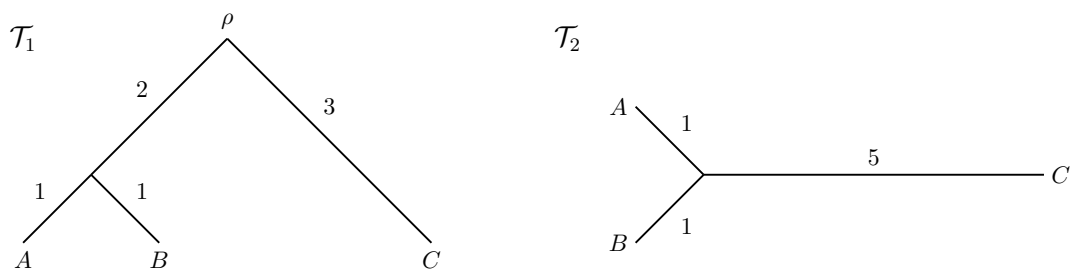


Fig. 24: Onw possible *Newick tree representation* for the rooted binary phylogentic X-tree $\mathcal{T}_1$ is $((A:1, B:1):2, C:3)$; and for the unrooted binary phylogenetic X-tree $\mathcal{T}_2$ one possible representation is $(A:1, B:1, C:5)$;.

Remark. Note that the Newick format of a given phylogenetic tree is not necessarily unique. Consider for example the rooted binary phylogenetic X-tree $\mathcal{T}_1$ depicted in Figure 24. We have the following Newick representations for $\mathcal{T}_1$:

- $((A:1, B:1):2, C:3);$
- $((B:1, A:1):2, C:3);$
- $(C:3, (A:1, B:1):2);$
- $(C:3, (B:1, A:1):2);$

In order to extend the Newick format to phylogenetic networks, two proposals were made: one represents a network as a single Newick string by splitting each reticulation node once for each parent, while the other represents it as a set of Newick strings by decomposing the network into a series of trees (cf. Cardona et al. [6]).

Both approaches have been referred to as the *extended Newick format*. We will shortly describe the two methods, although the focus will be on the first proposal.

First proposal

The main idea of the first approach is to modify the network so that it becomes a phylogenetic tree and then use the Newick tree format to describe this tree.

Formally, Algorithm 1 can be used to determine the Newick representation of a phylogenetic network (cf. Cardona et al. [6], Figure 25).

Remark. When reticulation nodes represent horizontal gene transfer events, we have to distinguish the HGT edge from the other edge directed into the reticulation node. 'This can easily be achieved by taking the target of the other edge as first replicate (the one that will carry the children of the [reticulation] node in the phylogenetic network) and the target of the reticulation edge as second replicate (the one that will become a terminal node) when splitting the [reticulation] node' (Cardona et al. [6]; cf. Figure 26).

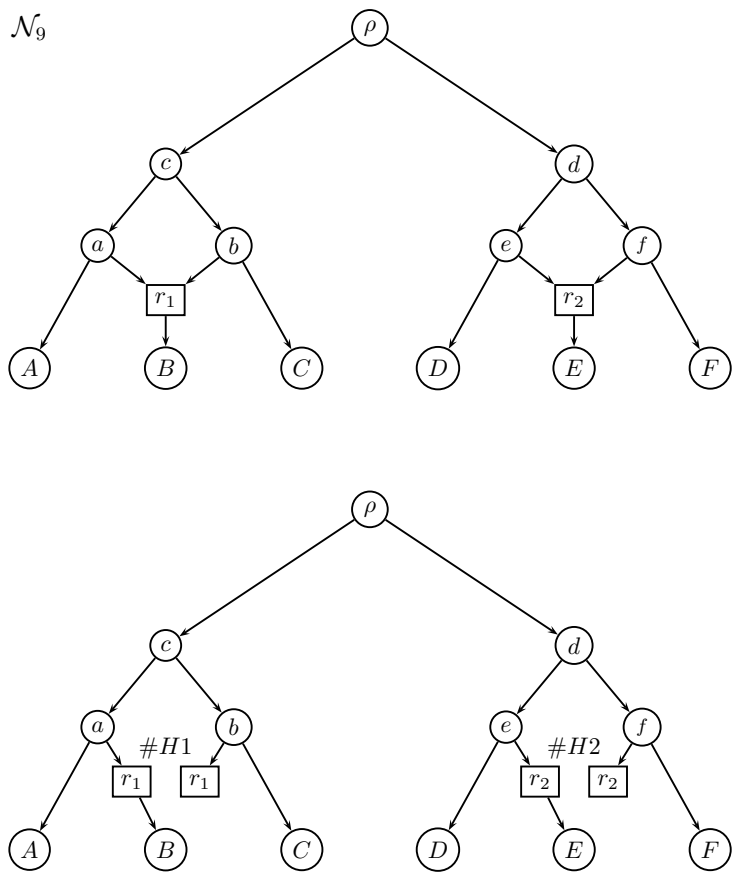


Fig. 25: The rooted phylogenetic network $\mathcal{N}_9$ with two reticulation nodes representing hybridization events can be transformed into a phylogenetic tree with two replicated nodes (r_1 and r_2). This leads to the *extended Newick* representation of $\mathcal{N}_9$ as

$$(((\mathsf{A},(\mathsf{B})\mathsf{r}_1\#\mathsf{H1})\mathsf{a},(\mathsf{r}_1\#\mathsf{H1},\mathsf{C})\mathsf{b})\mathsf{c},((\mathsf{D},(\mathsf{E})\mathsf{r}_2\#\mathsf{H2})\mathsf{e},(\mathsf{r}_2\#\mathsf{H2},\mathsf{F})\mathsf{f})\mathsf{d})\rho;$$

or, for short,

$$(((\mathsf{A},(\mathsf{B})\#\mathsf{H1}),(\#\mathsf{H1},\mathsf{C})),((\mathsf{D},(\mathsf{E})\#\mathsf{H2}),(\#\mathsf{H2},\mathsf{F})));.$$

Algorithm 1 Extendend Newick format (Phylogenetic network $\mathcal{N}$ with k reticulation nodes $r_1, \ldots, r_k$) (cf. Cardona et al. [6])

1: Split each reticulation node r_i with m parents $u_1, u_2, \ldots, u_m$ and children $v_1, v_2, \ldots, v$ in m different nodes, the first such copy with parent u_1 and children $v_1, v_2, \ldots, v$, and the remaining $m-1$ copies with a single parent $u_2, \ldots, u_k$, respectively, and no children.

2: Label each tree node with

$$[\text{label}][:\text{edge\_length}].$$

3: For all reticulation nodes r_i label the copies with

$$[\text{label}]\#[\text{type}]i[:\text{edge\_length}],$$

where label is an optional string providing a labeling of the node and type is an optional string indicating if the reticulation node represents a recombination (indicated by R), a hybridization (indicated by H) or a horizontal gene transfer (indicated by LGT) event. The obligatory integer i identifies the reticulation node r_i and edge_length is an optional number representing the edge length of the edge from the parent to the copy of r_i under consideration.

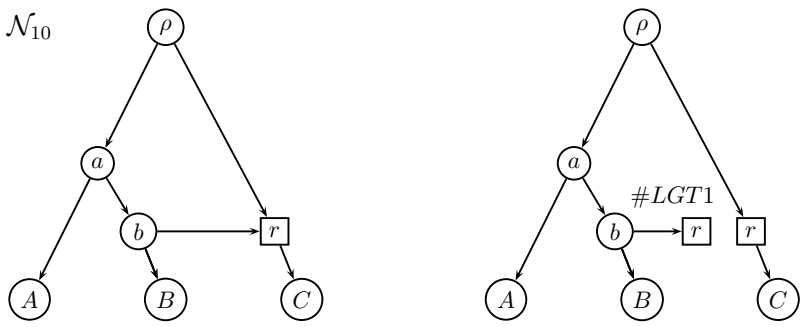


Fig. 26: The rooted phylogenetic network $\mathcal{N}_{10}$ with one reticulation node r representing a HGT (LGT) event can be transformed into a phylogenetic tree, where r is replicated. The target of the edge (ρ, r) is used as the first replicate (carrying the child C), while the target of the reticulation edge (b, r) is used as the second replicate, making the second copy of r become a leaf. This leads to the *extended Newick* representation of $\mathcal{N}_{10}$ as

$((\mathsf{A}, (\mathsf{B}, \mathsf{r}\#\mathsf{LGT1})\mathsf{b})\mathsf{a}, (\mathsf{C})\mathsf{r}\#\mathsf{LGT1})\rho;$

or, for short,

$((\mathsf{A}, (\mathsf{B}, \#\mathsf{LGT1})), (\mathsf{C})\#\mathsf{LGT1});$

(Figure taken from Cardona et al. [6] (slightly altered)).

Second proposal

The idea of the second approach is to 'break the network into a set of trees, and then represent the network as a collection of Newick representations of those trees' (Than et al. [32]). To be precise, a phylogenetic network with k hybrid nodes is represented as a forest of $k+1$ *multi-labeled*[11] phylogenetic trees. This is achieved by the following procedure (cf. Phy [2]):

1. Split each reticulation node r_i with m parents $u_1, u_2, \ldots, u_m$ and children $v_1, v_2, \ldots, v_l$ in $m+1$ different nodes.
2. Let each of the first m copies be a child of one of the nodes $u_1, \ldots, u_m$ (one for each) and have no children.
3. Let the last copy (the $m+1^{th}$ copy) of the reticulation node r_i have no parents and let it have the nodes $v_1, \ldots, v_l$ as its children.

The result is a forest, where each connected component is a multi-labeled phylogenetic tree, either rooted at the root of the network or at a copy of a reticulation node (Step 3).
The *extended Newick* representation of the network then consists of a string $n(t_1); n(t_2); \ldots n(t_{k+1})$, where $n(t_i)$ is the Newick representation of the tree t_i (cf. Figure 27).

Remark. A drawback of this approach is the fact that information about horizontal gene transfer events is lost, because the reticulation edge and the other edge coming into a reticulation node cannot be distinguished. Therefore Than et al. [32] suggest to include an explicit list of the horizontal gene transfer arrows in the representation of a HGT network (cf. Figure 28).

[11] Leaf labels may occur more than once (cf. Figure 28).

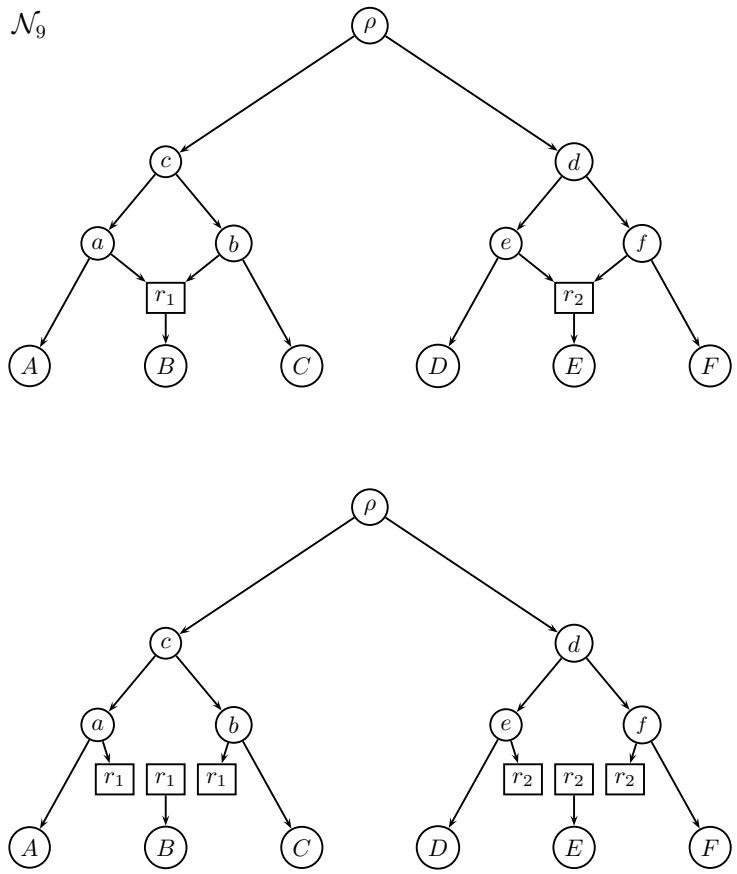


Fig. 27: The rooted phylogenetic network $\mathcal{N}_9$ with two reticulation nodes representing hybridization events can be decomposed into a series of three phylogenetic trees, namely $(B)r_1;$, $(E)r_2;$ and $(((A, r_1)a, (r_1, C)b)c, ((D, r_2)e, (r_2, F)f)d)\rho;$.
Note that the latter is formally a *multi-labeled phylogenetic tree*, because there are two leaves labeled with r_1 and two leaves labeled with r_2, thus, the leaf labels r_1 and r_2 occur more than once.
The *extended Newick* representation of $\mathcal{N}_9$ is

$(\mathsf{B})\mathsf{r}_1; (\mathsf{E})\mathsf{r}_2; (((\mathsf{A}, \mathsf{r}_1)\mathsf{a}, (\mathsf{r}_1, \mathsf{C})\mathsf{b})\mathsf{c}, ((\mathsf{D}, \mathsf{r}_2)\mathsf{e}, (\mathsf{r}_2, \mathsf{F})\mathsf{f})\mathsf{d})\rho;$

or, for short,

$(\mathsf{B}); (\mathsf{E}); (((\mathsf{A}, \mathsf{r}_1), (\mathsf{r}_1, \mathsf{C})), ((\mathsf{D}, \mathsf{r}_2), (\mathsf{r}_2, \mathsf{F})));.$

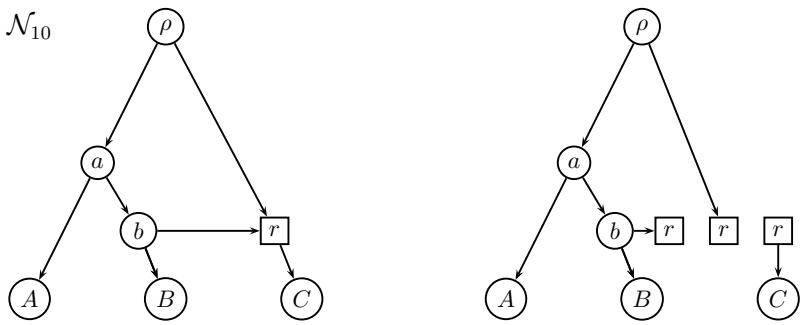


Fig. 28: The rooted phylogenetic network $\mathcal{N}_{10}$ with one reticulation node r representing a HGT (LGT) event can be decomposed into the phylogenetic tree $(C)r$; and the multi-labeled phylogenetic tree $((A,(B,r)b)a,r)\rho$;. To include information about the horizontal gene transfer event, the representation can be complemented by a list of the horizontal gene transfer arrows, i.e, $b \to r$. Thus, we retrieve the following representation of $\mathcal{N}_{10}$:

$(\mathsf{C})\mathsf{r}; ((\mathsf{A},(\mathsf{B},\mathsf{r})\mathsf{b})\mathsf{a},\mathsf{r})\rho; \mathsf{b} \to \mathsf{r}$

If the extended Newick representation is given *without* the horizontal gene transfer arrow $b \to r$, we cannot distinguish between the horizontal gene transfer edge and the other edge directed into r (Figure in dependence on Cardona et al. [6]).

Hybridization probabilities Note that neither version of the extended Newick format for phylogenetic networks allows for the inclusion of hybridization probabilities.

However, there have been (unpublished) suggestions for a so-called *Rich Newick Format*, which is based on the extended Newick Format and enables the inclusion of hybridization probabilities and support values (cf. Ric [3]). This is achieved by extending the information associated with a node (either tree node or reticulation node) in the network by additional labeling:

- Tree node:

  [label][:edge_length][:support][:probability]

- Reticulation node:

  [label]#[type]i[:edge_length][:support][:probability]

The *Rich Newick Format* is apparently associated with the software package *PhyloNet* (Than et al. [32]) and seems like a promising approach towards the representation of phylogenetic networks.

However, it does not seem to be in widespread use elsewhere. In particular, the *BioPerl* toolkit (Stajich [30]), on which the implementation is based in this thesis, uses

the extended Newick format rather than the Rich Newick format for the representation of phylogenetic networks.

6.2. Computation of generalized phylogenetic diversity and biodiversity indices

The implementation in this thesis is based on the *BioPerl* toolkit (Stajich [30]), in particular on the Perl module `Bio::PhyloNetwork` (Cardona et al. [7]), which requires the extended Newick format representation of a phylogenetic network.[12] However, the extended Newick format has the drawback that hybridization probabilities cannot be included in the representation of a phylogenetic network (cf. page 102).

Therefore we are not able to implement all measures of generalized *phylogenetic diversity* and all generalized biodiversity indices introduced in previous chapters, but will focus on the approaches independent of hybridization probabilities.

In the following we show how to compute and implement the following measures of generalized *phylogenetic diversity*

- PND,
- $PD^{*}_{\mathsf{T}(\mathcal{N})}$ with $* \in \{\min, \max, \sum, \varnothing\}$,
- PD^{LSA}

and the following generalized biodiversity indices

- $FP^{*}_{\mathsf{T}(\mathcal{N})}$, $SV^{*}_{\mathsf{T}(\mathcal{N})}$, $\widetilde{SV}^{*}_{\mathsf{T}(\mathcal{N})}$ and $\widehat{SV}^{*}_{\mathsf{T}(\mathcal{N})}$ with $* \in \{\min, \max, \sum, \varnothing\}$,
- FP^{LSA}, SV^{LSA}, $\widetilde{SV}^{LSA}$ and $\widehat{SV}^{LSA}$,
- $SV_{\mathcal{PD}}$ and $\widetilde{SV}_{\mathcal{PD}}$ with $\mathcal{PD} \in \{PND, PD^{\min}_{\mathsf{T}(\mathcal{N})}, PD^{\max}_{\mathsf{T}(\mathcal{N})}, PD^{\sum}_{\mathsf{T}(\mathcal{N})}, PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}, PD^{LSA}\}$.

Before going into the details of the computation of each individual measure or index, we will describe some general methods needed for their computation.

[12] The Perl module can take both versions (i.e. proposal 1 and proposal 2) of the extended Newick format as input, but uses the first version internally. In this thesis we also use the first proposal of the extended Newick format.

6.2.1. Basic principles

Computation of the set of embedded trees In order to calculate the (multi)set $\mathsf{T}(\mathcal{N})$ of phylogenetic X-trees displayed by a phylogenetic network $\mathcal{N}$ on X, we use the method `explode` provided by the Perl module `Bio::PhyloNetwork` (Cardona et al. [7]). However, this method does not return the (multi)set $\mathsf{T}(\mathcal{N})$ of phylogenetic X-trees, but the extended (multi)set $\overline{\mathsf{T}(\mathcal{N})}$ of all trees displayed by $\mathcal{N}$. In a second step, we therefore discard all trees that are not phylogenetic X-trees (cf. Algorithm 2) .

Algorithm 2 Set of embedded phylogenetic X-trees

Input: phylogenetic network $\mathcal{N}$ on taxon set X
Output: set of phylogenetic X-trees displayed by $\mathcal{N}$

1: $\mathsf{T}(\mathcal{N}) := \emptyset$; ▷ (Multi)set of embedded phylogenetic X-trees.
2: $\overline{\mathsf{T}(\mathcal{N})} := \mathcal{N} \rightarrow$ `explode()`; ▷ (Multi)set of all embedded trees.
3: **for all** $\mathcal{T} \in \overline{\mathsf{T}(\mathcal{N})}$ **do**
4: $X' :=$ leaf set of $\mathcal{T}$;
5: **if** $X' = X$, i.e. if $\mathcal{T}$ and $\mathcal{N}$ have the same leaf set **then**
6: $\mathsf{T}(\mathcal{N}) := \mathsf{T}(\mathcal{N}) \cup \{\mathcal{T}\}$, i.e. add $\mathcal{T}$ to $\mathsf{T}(\mathcal{N})$;
7: **end if**
8: **end for**
9: **return** $\mathsf{T}(\mathcal{N})$;

Computation of the *LSA tree* associated with a phylogenetic network The computation of the *LSA tree* associated with a phylogenetic network $\mathcal{N}$ requires three major steps:

- Finding the lowest stable ancestor for all reticulation nodes.
- Calculating the path lengths of all paths between the lowest stable ancestor $lsa(r)$ of a reticulation node r and r.
- Constructing the topology of the *LSA tree* and assigning branch lengths to it.

In order to compute the lowest stable ancestor for all reticulation nodes (or, to be precise, for all nodes), e.g. the Lengauer Tarjan Algorithm (Lengauer and Tarjan [24]) can be used. However, we have not implemented this algorithm, but use a more intuitive (although more time-complex) approach.
We first compute the set $SA(v)$ of all *stable ancestors* for all nodes v in $\mathcal{N}$ (cf. Definition

16) by the following equations

$$SA(\rho) = \{\rho\}, \tag{6.1}$$

$$SA(v) = \Big(\bigcap_{\substack{w:w \text{ is} \\ \text{ancestor of } v}} SA(w) \Big) \cup \{v\}, \tag{6.2}$$

where ρ is the root of $\mathcal{N}$. We then consider the subset of reticulation nodes of $\mathcal{N}$ and set the *lowest stable ancestor* of a reticulation node r as the node $lsa(r) \in SA(r) \setminus r$ that is furthest away from the root ρ. In order to find the node $lsa(r) \in SA(r) \setminus r$ that is furthest away from the root ρ, we sort the nodes of the network *topologically*, i.e. we produce a linear ordering of the nodes such that for every directed edge $e = (u, v)$, the node u comes before the node v in the ordering.
(If a graph contains cycles, no linear ordering of the nodes is possible. Phylogenetic networks, however, are *directed acyclic graphs* and thus, it is always possible to find a linear ordering of its nodes, but this ordering is not necessarily unique (cf. Figure 29). A (not necessarily unique) topological ordering of the nodes in $V(\mathcal{N})$ can for example be found by repeating the following steps until the graph is empty:

1. Choose any node $v \in V(\mathcal{N})$ with indegree 0.
2. Add v to the topological ordering.
3. Delete v and its incident edges from $\mathcal{N}$.)

Let O^{top} be a linear ordering of the nodes of $\mathcal{N}$. Then the lowest stable ancestor $lsa(r) \in SA(r) \setminus r$ of a reticulation node r is the last node in the topological ordering $O^{top} \restriction SA(r) \setminus r$ restricted to the nodes in $SA(r) \setminus r$ (cf. Figure 29).

Note that we have to compute the sets $SA(v)$ for all nodes v in $\mathcal{N}$ in order to retrieve the set $SA(r)$ of stable ancestors of a reticulation node r and subsequently, its lowest stable ancestor $lsa(r)$.

For the calculation of all paths between the lowest stable ancestor of a reticulation node and the reticulation node we use a recursive search, where we successively build up the paths. However, in order to avoid unnecessary calculations (e.g. 'dead ends'), we reverse the direction of all edges in $\mathcal{N}$ and successively calculate all paths between the reticulation node and its lowest stable ancestor. This assures that no unnecessary paths are calculated, because by definition the lowest stable ancestor of a node r lies on all paths from r to the root, in particular it lies on all paths starting from r (if the direction of all edges is reversed).

We then construct the topology of the *LSA tree* according to its definition (cf. Definition 17), i.e. for each reticulation node r in $\mathcal{N}$, we remove all edges directed into r and

add a new edge $e = (lsa(r), r)$ from the lowest stable ancestor of r into r. We thereby assign the average length of a path between $lsa(r)$ and r to the new edge $e = (lsa(r), r)$. All other edges (i.e. the tree edges of $\mathcal{N}$) keep their branch lengths, respectively. We then repeatedly remove all unlabeled leaves and nodes with in- and outdegree 1, until no further such removal is possible. Thereby formerly distinct edges may be melted into one edge, in which case their edge lengths are added to yield the edge length of the new edge (cf. Algorithm 3).

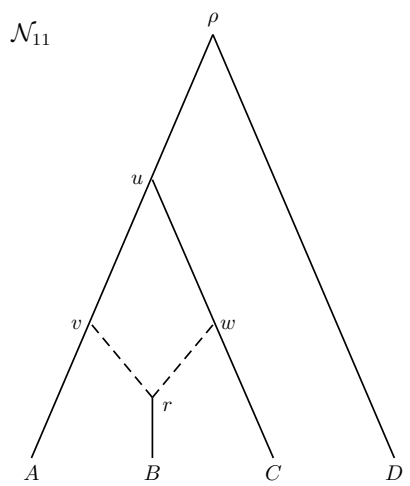


Fig. 29: Rooted binary phylogenetic network $\mathcal{N}_{11}$ on $X = \{A, B, C, D\}$. A topological ordering of the nodes of $\mathcal{N}_{11}$ is, for example, $O_1^{top} = \rho$, D, u, v, A, w, C, r, B (The topological ordering of $\mathcal{N}_{11}$ is not unique. $O_2^{top} = \rho$, u, D, v, A, w, C, r, B is also a valid linear ordering and there are many more.)
For the reticulation node r, we have $SA(r) \setminus r = \{\rho, u\}$. In order to find the node $lsa(r) \in SA(r) \setminus r$ that is furthest away from the root ρ, we consider a topological ordering of $\mathcal{N}_{11}$ (e.g. O_1^{top}) restricted to the nodes in $SA(r) \setminus r$. We have $O_1^{top} \restriction SA(r) \setminus r = \rho$, u and thus u is the lowest stable ancestor of r, because it occupies the last position in the restricted linear ordering.

Algorithm 3 LSA Tree associated with a phylogenetic network

Input: phylogenetic network $\mathcal{N}$ on taxon set X
Output: *LSA tree* associated with $\mathcal{N}$

1: **Compute the lowest stable ancestors for all reticulation nodes**
2: $V \coloneqq$ nodes of $\mathcal{N}$ sorted topologically; ▷ We have a linear ordering of the nodes of $\mathcal{N}$.
3: $SA(\rho) \coloneqq \{\rho\}$; ▷ Initialize the set of stable ancestors of the root.
4: **for all** $v \in V \setminus \rho$ **do**
5: **if** v is a tree node **then**
6: $SA(v) \coloneqq SA(p) \cup \{v\}$, where p is the parent of v; ▷ If v is a tree node, it has only one predecessor, its parent, p.
7: **else**
8: $SA(v) \coloneqq \Big(\bigcap_{w:w \text{ is ancestor of } v} SA(w) \Big) \cup \{v\}$; ▷ Compute the set of stable ancestors for a reticulation node.
9: $lsa(v) \coloneqq u$, where u is in $SA(v) \setminus \{v\}$ and is furthest away from the root ρ;
10: **end if**
11: **end for**
12:
13: **Calculate the average path length between a reticulation node and its lowest stable ancestor**
14: $\mathcal{N}' \coloneqq \mathcal{N}^\top$; ▷ Reverse the direction of all edges.
15: **for all** reticulation nodes r of $\mathcal{N}$ **do**
16: Calculate the set $\mathcal{P}_r$ of all paths between r and $lsa(r)$ in $\mathcal{N}'$;
17: Calculate the average path length as $\lambda_r \coloneqq \frac{1}{|\mathcal{P}_r|} \sum_{P \in \mathcal{P}_r} length(P)$;
18: **end for**
19:
20: **Construct the *LSA tree*** introduction to algorithms
21: $\mathcal{T}_{LSA}(\mathcal{N}) \coloneqq \mathcal{N}$; ▷ Initialize the *LSA tree* as $\mathcal{N}$.
22: **for all** reticulation nodes r of $\mathcal{N}$ **do**
23: Remove all edges directed into r from $\mathcal{T}_{LSA}(\mathcal{N})$;
24: Add a new edge $e = (lsa(r), r)$ to $\mathcal{T}_{LSA}(\mathcal{N})$ and use λ_r as its edge length;
25: **end for**
26: Remove all unlabeled leaves from $\mathcal{T}_{LSA}(\mathcal{N})$;
27: Remove all nodes with in- and outdegree 1 from $\mathcal{T}_{LSA}(\mathcal{N})$;
28:
29: **return** $\mathcal{T}_{LSA}(\mathcal{N})$;

Computation of the phylogenetic diversity of a subset of taxa In order to calculate the *phylogenetic diversity* of a subset $S \subseteq X$ of taxa of a rooted phylogenetic X-tree $\mathcal{T}$,[13] we compute the (unique) path from the root ρ of $\mathcal{T}$ to a taxon $a \in S$ for all taxa. The *phylogenetic diversity* of S then calculates as the sum of edge lengths of these paths, where each edge is only taken into account once (even though it may be part of several paths) (cf. Algorithm 4).

Algorithm 4 Phylogenetic diversity

Input: rooted phylogenetic X-tree $\mathcal{T}$, subset $S \subseteq X$ of taxa
Output: phylogenetic diversity $PD(S)$

1: $E_S := \emptyset$; ▷ Set to store the edges of $\mathcal{T}$ connecting the taxa in S and the root.
2: **for all** $a \in S$ **do**
3: Compute the path P_a from the root ρ to a;
4: $E_a :=$ edges of P_a; ▷ Edges on the path P_a from the root to taxon a.
5: $E_S := E_S \cup E_a$; ▷ Add the edges of P_a to the set E_S; as E_S is a set (not a multiset), each edge is only taken into account once.
6: **end for**
7: $PD(S) := \sum_{e \in E_S} length(e)$; ▷ Add the edge lengths of all edges in E_S to obtain $PD(S)$.
8: **return** $PD(S)$;

Computation of the Fair Proportion Index and the different versions of the Shapley Value The computation of the different biodiversity indices is based on the algorithms presented in Wicke [36], and for the implementation of the indices we partly use source code from `FairShapley`, a software tool introduced by Wicke and Fischer [37]. However, for the sake of completeness, we shortly describe the relevant algorithms.

The Fair Proportion Index The *Fair Proportion Index* for the taxa of a rooted phylogenetic X-tree $\mathcal{T}$ can be computed easily according to its definition (cf. Definition 5). We first loop over the edges of $\mathcal{T}$ and calculate the contribution $c(e)$ of an edge e to the *Fair Proportion Index* by dividing the edge length λ_e of e by the number of its descendent leaves D_e, respectively. In a second step we loop over the leaves of $\mathcal{T}$ and calculate the *Fair Proportion Index* for each leaf by summing up the contributions $c(e)$ for all edges e on the path from the root to the leaf (cf. Algorithm 5).

[13] We only consider rooted phylogenetic X-trees here, because the *phylogenetic diversity* of unrooted phylogenetic X-trees is not needed for further calculations.

Algorithm 5 Fair Proportion Index

Input: rooted phylogenetic X-tree $\mathcal{T}$
Output: Fair Proportion Indices for all taxa $a \in X$

1: **for all** edges e of $\mathcal{T}$ **do**
2: $\lambda_e := length(e)$;
3: $D_e :=$ number of descendent leaves of e ;
4: $c(e) := \frac{\lambda_e}{D_e}$; ▷ Contribution of an edge e to the Fair Proportion Index.
5: **end for**
6: **for all** taxa $a \in X$ of $\mathcal{T}$ **do**
7: $FP(a) := \sum_e c(e)$, where the sum runs over all edges e on the path from the root to a;
8: **end for**
9: **return** $FP(a)$ for all $a \in X$;

The original and the modified Shapley Value For rooted phylogenetic X-trees the *Fair Proportion Index* and the *original Shapley Value* coincide for all taxa $a \in X$ (cf. Fuchs and Jin [15]). Thus, we use Algorithm 5 for the calculation of the *original Shapley Value* (cf. Algorithm 6).

Algorithm 6 Original Shapley Value

Input: rooted phylogenetic X-tree $\mathcal{T}$
Output: original Shapley Values for all taxa $a \in X$

1: **for all** taxa $a \in X$ of $\mathcal{T}$ **do**
2: Use Algorithm 5 to calculate $FP(a)$;
3: Set $SV(a) := FP(a)$;
4: **end for**
5: **return** $SV(a)$ for all $a \in X$;

The *modified Shapley Value* of a taxon $a \in X$, on the other hand, can be derived from the *Fair Proportion Index* of a as

$$\begin{aligned}\widetilde{SV}(a) &= SV(a) - \frac{PD(\{a\})}{n} \\ &= FP(a) - \frac{PD(\{a\})}{n},\end{aligned}$$

where $PD(\{a\})$ is the *phylogenetic diversity* of $\{a\}$ and $n = |X|$ is the number of taxa of $\mathcal{T}$ (cf. Proposition 1).

Thus, in order to compute the *modified Shapley Value* for a taxon a of a rooted phylogenetic X-tree, we first calculate the *Fair Proportion Index* of a and then subtract the term $\frac{PD(\{a\})}{n}$ (cf. Algorithm 7).

Algorithm 7 Modified Shapley Value

Input: rooted phylogenetic X-tree $\mathcal{T}$
Output: modified Shapley Values for all taxa $a \in X$

1: $n := |X|$;
2: **for all** taxa $a \in X$ of $\mathcal{T}$ **do**
3: Use Algorithm 5 to calculate $FP(a)$;
4: Use Algorithm 4 to calculate $PD(\{a\})$;
5: $\widetilde{SV}(a) := FP(a) - \frac{PD(\{a\})}{n}$;
6: **end for**
7: **return** $\widetilde{SV}(a)$ for all $a \in X$;

In some cases (e.g. for certain versions of the *generalized Shapley Value*), it is necessary to calculate the *Shapley Value* according to its definition. In this case we require the *phylogenetic diversity* of all subsets $S \subseteq X$ of taxa, but not the phylogenetic X-tree or the phylogenetic network on X, respectively (cf. Algorithm 8, Algorithm 9).

Algorithm 8 Original Shapley Value from PD

Input: all subsets $S \subseteq X$ of taxa and their phylogenetic diversity $PD(S)$
Output: original Shapley Values for all taxa $a \in X$

1: $n := |X|$;
2: **for all** taxa $a \in X$ **do**
3:

$$SV(a) = \frac{1}{n!} \sum_{\substack{S \subseteq X \\ a \in S}} \Big((|S| - 1)!(n - |S|)!(PD(S) - PD(S \setminus \{a\})) \Big);$$

▷ Sum runs over all subsets $S \subseteq X$ containing a.

4: **end for**
5: **return** $SV(a)$ for all $a \in X$;

Algorithm 9 Modified Shapley Value from PD

Input: all subsets $S \subseteq X$ of taxa and their phylogenetic diversity $PD(S)$
Output: modified Shapley Values for all taxa $a \in X$

1: $n := |X|$;
2: **for all** taxa $a \in X$ **do**
3:
$$\widetilde{SV}(a) = \frac{1}{n!} \sum_{\substack{S \subseteq X : a \in S \\ |S| \geq 2}} \Big((|S|-1)!(n-|S|)!(PD(S) - PD(S \setminus \{a\})) \Big);$$
▷ Sum runs over all subsets $S \subseteq X$ containing a and at least one other taxon.
4: **end for**
5: **return** $\widetilde{SV}(a)$ for all $a \in X$;

The unrooted rooted Shapley Value The *original Shapley Value* of unrooted phylogenetic X-trees and thus, the *unrooted rooted Shapley Value* of rooted phylogenetic X-trees can be calculated by considering so-called X-splits induced by the edges of the trees (cf. Haake et al. [16], Wicke [36]), which we will define in the following. This approach is less complex than the calculation according to the definition of the *Shapley Value* and therefore requires less computation time.

Definition 40 (X-Split; induced X-Split)**.**

1. An X-*split* $\sigma = A|B$ is a bipartition of a set X into two, non-empty, disjoint sets A and B, i.e. $X = A \cup B$, $A \cap B = \emptyset$ and $A, B \neq \emptyset$.

2. Let $\mathcal{T}$ be a phylogenetic X-tree and let e be an edge of $\mathcal{T}$. Then $\mathcal{T} \setminus e$ consists of two connected components (subtrees) $\mathcal{T}_1$ and $\mathcal{T}_2$. Let $X_i \subseteq X$ be the subset of taxa belonging to $\mathcal{T}_i$ $(i = 1, 2)$. Then $\sigma_e := X_1|X_2$ is called the X-*split induced by* e.

Idea (cf. Haake et al. [16])**.** Let $\mathcal{T}_u$ be an unrooted phylogenetic X-tree. Let E denote the set of edges of $\mathcal{T}_u$. For $a \in X$ and $e \in E$, the removal of the edge e splits $\mathcal{T}_u$ into two subtrees. Let $C(a, e)$ denote the set of leaves in the subtree that contains a and let $F(a, e)$ denote the set of leaves in the other subtree, that is 'far' from a.
Then the *original Shapley Value* for a taxon $a \in X$ can be calculated as

$$\begin{aligned} SV(a) &= \frac{1}{n!} \sum_{\substack{S \subseteq X \\ a \in S}} \Big((|S|-1)!(n-|S|)!(PD(S) - PD(S \setminus \{a\})) \Big) \\ &= \sum_{e \in E} \lambda_e \frac{|F(a,e)|}{n|C(a,e)|}, \end{aligned} \tag{6.3}$$

where the sum runs over all edges of $\mathcal{T}_u$ with edge lengths λ_e and $n = |X|$ is the total number of taxa. The idea of Equation (6.3) is to count the number of times the weight λ_e of a given edge e is in the marginal contribution $PD(S) - PD(S \setminus \{a\})$ of a for coalitions of size $|S|$. This is the case, if the other $|S| - 1$ members of the coalition are in the subtree that is *far* from a (if the edge e is removed). For details of the proof, see Haake et al. [16].

Example 21. Consider the unrooted phylogenetic X-tree $\mathcal{T}_2$ depicted in Figure 1. We can calculate the *original Shapley Value* of A as

$$\begin{aligned} SV_{\mathcal{T}_2}(A) &= 1 \cdot \frac{2}{3 \cdot 1} + 1\frac{1}{3 \cdot 2} + 5 \cdot \frac{1}{3 \cdot 2} \\ &= \frac{2}{3} + \frac{1}{6} + \frac{5}{6} \\ &= \frac{5}{3}. \end{aligned}$$

Note that this calculation is less complex than the calculation of $SV_{\mathcal{T}_2}(A)$ according to its definition (cf. Example 3).

In order to compute the *unrooted rooted Shapley Value* for the taxa of a rooted phylogenetic X-tree $\mathcal{T}$, we unroot $\mathcal{T}$ by suppressing the root node ρ. We then loop over all edges e of $\mathcal{T}$ and consider the X-split induced by e, i.e. we consider the bipartition of the taxon set X into two subsets $X_1(e)$ and $X_2(e)$ and then use Equation (6.3) to calculate the *original Shapley Value* of a taxon $a \in X$ (cf. Algorithm 10). Note that

$$|F(a,e)| = \begin{cases} |X_1(e)| & \text{if } a \in X_2(e) \\ |X_2(e)| & \text{if } a \in X_1(e) \end{cases} \qquad \text{and} \qquad |C(a,e)| = \begin{cases} |X_1(e)| & \text{if } a \in X_1(e) \\ |X_2(e)| & \text{if } a \in X_2(e) \end{cases}.$$

Based on these general methods we can now describe how the individual measures of generalized *phylogenetic diversity* and generalized biodiversity indices are computed.

Algorithm 10 Unrooted rooted Shapley Value

Input: rooted phylogenetic X-tree $\mathcal{T}_r$
Output: unrooted rooted Shapley Values for all taxa $a \in X$

1: Unroot $\mathcal{T}_r$ to retrieve an unrooted X-tree $\mathcal{T}_u$;
2: $n \coloneqq |X|$;
3: **for all** edges e in $\mathcal{T}_u$ **do**
4: Calculate $\sigma_e = X_1(e)|X_2(e)$; ▷ Split of the taxon set X induced by the removal of edge e.
5: **end for**
6: **for all** taxa $a \in X$ **do**
7: $\widehat{SV}(a) \coloneqq 0$; ▷ Initialize $\widehat{SV}(a)$.
8: **for all** edges e in $\mathcal{T}_u$ **do**
9: **if** $a \in X_1(e)$ **then**
10: $|F(a,e)| \coloneqq |X_2(e)|$ and $|C(a,e)| \coloneqq |X_1(e)|$;
11: **else**
12: $|F(a,e)| \coloneqq |X_1(e)|$ and $|C(a,e)| \coloneqq |X_2(e)|$;
13: **end if**
14: $\widehat{SV}(a) \coloneqq \widehat{SV}(a) + \lambda_e \frac{|F(a,e)|}{n \cdot |C(l,e)|}$, where λ_e is the length of edge e;
15: **end for**
16: **end for**
17: **return** $\widehat{SV}(a)$ for all taxa $a \in X$;

6.2.2. Generalized phylogenetic diversity

Phylogenetic net diversity In order to calculate the *phylogenetic net diversity* of a subset $S \subseteq X$ of taxa of a phylogenetic network $\mathcal{N}$ on X, we consider the (multi)set $\overline{\mathsf{T}(\mathcal{N})}$ of all trees (not necessarily phylogenetic X-trees) displayed by $\mathcal{N}$ and use the relationship (cf. second remark on page 53 ff.)

$$PND(S) = PD^{\min}_{\overline{\mathsf{T}(\mathcal{N})}}(S)$$

for the computation (cf. Algorithm 11).

Algorithm 11 Phylogenetic net diversity

Input: rooted phylogenetic network $\mathcal{N}$ on some taxon set X, subset $S \subseteq X$ of taxa
Output: phylogenetic net diversity $PND(S)$

1: $\overline{\mathsf{T}(\mathcal{N})} \coloneqq \mathcal{N} \to$ `explode()`; ▷ (Multi)set of all embedded trees.
2: **for all** trees $\mathcal{T}$ in $\overline{\mathsf{T}(\mathcal{N})}$ **do**
3: Use Algorithm 4 to calculate the phylogenetic diversity $PD_{\mathcal{T}}(S)$ of S;
4: **end for**
5: $PND(S) \coloneqq \min\limits_{\mathcal{T} \in \overline{\mathsf{T}(\mathcal{N})}} PD_{\mathcal{T}}(S)$;
6: **return** $PND(S)$;

Embedded phylogenetic diversity The computation of the *embedded phylogenetic diversity* is similar to the computation of the *phylogenetic net diversity*, but we now consider the (multi)set $\mathsf{T}(\mathcal{N})$ of phylogenetic X-trees displayed by $\mathcal{N}$. We then calculate the *phylogenetic diversity* of $S \subseteq X$ for all embedded phylogenetic X-trees and derive the *embedded phylogenetic diversity* as the minimum, maximum, sum or average thereof (cf. Algorithm 12).

Algorithm 12 Emebdded phylogenetic diversity

Input: rooted phylogenetic network $\mathcal{N}$ on some taxon set X, subset $S \subseteq X$ of taxa
Output: embedded phylogenetic diversity $PD^{*}_{\mathsf{T}(\mathcal{N})}(S)$ with $* \in \{\min, \max, \sum, \varnothing\}$

1: Use Algorithm 2 to retrieve the (multi)set $\mathsf{T}(\mathcal{N})$ of phylogenetic X-trees displayed by $\mathcal{N}$;
2: **for all** trees $\mathcal{T}$ in $\mathsf{T}(\mathcal{N})$ **do**
3: Use Algorithm 4 to calculate the phylogenetic diversity $PD_{\mathcal{T}}(S)$ of S;
4: **end for**
5: $PD^{\min}_{\mathsf{T}(\mathcal{N})}(S) := \min\limits_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} PD_{\mathcal{T}}(S)$;
6: $PD^{\max}_{\mathsf{T}(\mathcal{N})}(S) := \max\limits_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} PD_{\mathcal{T}}(S)$;
7: $PD^{\sum}_{\mathsf{T}(\mathcal{N})}(S) := \sum\limits_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} PD_{\mathcal{T}}(S)$;
8: $PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}(S) := \frac{1}{|\mathsf{T}(\mathcal{N})|} \sum\limits_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} PD_{\mathcal{T}}(S)$;
9: **return** $PD^{*}_{\mathsf{T}(\mathcal{N})}(S)$ with $* \in \{\min, \max, \sum, \varnothing\}$;

LSA associated phylogenetic diversity In order to calculate the *LSA associated phylogenetic diversity* for a subset $S \subseteq X$ of taxa of a phylogenetic network $\mathcal{N}$ on X, we construct the *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N})$ associated with $\mathcal{N}$ and calculate the *phylogenetic diversity* based on this tree (cf. Algorithm 13).

Algorithm 13 LSA associated phylogenetic diversity

Input: rooted phylogenetic network $\mathcal{N}$ on some taxon set X, subset $S \subseteq X$ of taxa
Output: LSA associated phylogenetic diversity $PD^{LSA}(S)$

1: Use Algorithm 3 to construct the *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N})$ associated with $\mathcal{N}$;
2: Use Algorithm 4 to calculate the phylogenetic diversity of S for $\mathcal{T}_{LSA}(\mathcal{N})$ and set

$$PD^{LSA}(S) := PD_{\mathcal{T}_{LSA}(\mathcal{N})}(S);$$

3: **return** $PD^{LSA}(S)$;

6.2.3. Generalized biodiversity indices

Embedded biodiversity indices For the calculation of the *embedded Fair Proportion Index* or the different versions of the *embedded Shapley Value* for the taxa of a phylogenetic network $\mathcal{N}$, we consider the (multi)set $\mathsf{T}(\mathcal{N})$ of phylogenetic X-trees displayed by $\mathcal{N}$. We then calculate the biodiversity index under consideration for all embedded X-trees and derive the embedded biodiversity index as the minimum, maximum, sum or average thereof. As the procedure is the same for the *embedded Fair Proportion Index* and all versions of the *embedded Shapley Value*, we only consider the *embedded original Shapley Value* here (cf. Algorithm 14).

Algorithm 14 Embedded original Shapley Value

Input: rooted phylogenetic network $\mathcal{N}$ on some taxon set X
Output: embedded original Shapley $SV^*_{\mathsf{T}(\mathcal{N})}(a)$ with $* \in \{\min, \max, \sum, \varnothing\}$ for all taxa $a \in X$

1: Use Algorithm 2 to retrieve the (multi)set $\mathsf{T}(\mathcal{N})$ of phylogenetic X-trees displayed by $\mathcal{N}$;
2: **for all** taxa $a \in X$ **do**
3: **for all** trees $\mathcal{T}$ in $\mathsf{T}(\mathcal{N})$ **do**
4: Use Algorithm 6 to calculate the original Shapley Value $SV(a)$;
5: **end for**
6: $SV^{\min}_{\mathsf{T}(\mathcal{N})}(a) := \min\limits_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} SV_{\mathcal{T}}(a)$;
7: $SV^{\max}_{\mathsf{T}(\mathcal{N})}(a) := \max\limits_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} SV_{\mathcal{T}}(a)$;
8: $PD^{\sum}_{\mathsf{T}(\mathcal{N})}(a) := \sum\limits_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} SV_{\mathcal{T}}(a)$;
9: $PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}(a) := \frac{1}{|\mathsf{T}(\mathcal{N})|} \sum\limits_{\mathcal{T} \in \mathsf{T}(\mathcal{N})} SV_{\mathcal{T}}(a)$;
10: **end for**
11: **return** $SV^*_{\mathsf{T}(\mathcal{N})}(a)$ with $* \in \{\min, \max, \sum, \varnothing\}$ for all $a \in X$;

LSA associated biodiversity indices In order to calculate the *LSA associated Fair Proportion Index* or the different versions of the *LSA associated Shapley Value* for the taxa of a phylogenetic network $\mathcal{N}$, we construct the *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N})$ associated with $\mathcal{N}$ and calculate the indices based on this tree. Again, we only consider the *LSA associated original Shapley Value* here, but the procedure is the same for all other LSA associated biodiversity indices (cf. Algorithm 15).

Algorithm 15 LSA associated original Shapley Value

Input: rooted phylogenetic network $\mathcal{N}$ on some taxon set X
Output: LSA associated original Shapley Value $SV^{LSA}(a)$ for all taxa $a \in X$

1: Use Algorithm 3 to construct the *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N})$ associated with $\mathcal{N}$;
2: **for all** taxa $a \in X$ **do**
3: Use Algorithm 6 to calculate the original Shapley Value of a for $\mathcal{T}_{LSA}(\mathcal{N})$ and set

$$SV^{LSA}(a) := SV_{\mathcal{T}_{LSA}}(a);$$

4: **end for**
5: **return** $SV^{LSA}(a)$ for all $a \in X$;

Generalized Shapley Value For the *generalized original Shapley Value* $SV_{\mathcal{PD}}$ and the *generalized modified Shapley Value* $\widetilde{SV}_{\mathcal{PD}}$, the method of computation depends on the measure $\mathcal{PD} \in \{PND, PD^{\min}_{\mathsf{T}(\mathcal{N})}, PD^{\max}_{\mathsf{T}(\mathcal{N})}, PD^{\sum}_{\mathsf{T}(\mathcal{N})}, PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}, PD^{LSA}(S)\}$ of generalized *phylogenetic diversity* used (cf. Algorithm 16). Again, we only consider the *generalized original Shapley Value*, but the same holds for the *generalized modified Shapley Value*.

- If $\mathcal{PD} \in \{PND, PD^{\min}_{\mathsf{T}(\mathcal{N})}, PD^{\max}_{\mathsf{T}(\mathcal{N})}\}$ we calculate the *generalized original Shapley Value* according to its definition, i.e. we use Algorithm 8.

- If $\mathcal{PD} \in \{PD^{\sum}_{\mathsf{T}(\mathcal{N})}, PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}\}$ we consider the (multi)set of embedded trees and use the following relationships

$$SV_{PD^{\sum}_{\mathsf{T}(\mathcal{N})}}(a) = SV^{\sum}_{\mathsf{T}(\mathcal{N})}(a),$$
$$SV_{PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}}(a) = SV^{\varnothing}_{\mathsf{T}(\mathcal{N})}(a)$$

  from Proposition 6.

- If $\mathcal{PD} = PD^{LSA}$ we calculate the *LSA associated original Shapley Value*.

Algorithm 16 Generalized Shapley Value

Input: rooted phylogenetic network $\mathcal{N}$ on some taxon set X

Output: generalized original Shapley Value $SV_{\mathcal{PD}}(a)$ with $\mathcal{PD} \in \{PND, PD^{\min}_{\mathsf{T}(\mathcal{N})}, PD^{\max}_{\mathsf{T}(\mathcal{N})}, PD^{\sum}_{\mathsf{T}(\mathcal{N})}, PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}, PD^{LSA}\}$ for all taxa $a \in X$

1: **Case 1:** $\mathcal{PD} \in \{PND, PD^{\min}_{\mathsf{T}(\mathcal{N})}, PD^{\max}_{\mathsf{T}(\mathcal{N})}\}$

2: Use Algorithm 11 to calculate the phylogenetic net diversity $PND(S)$ for all subsets $S \subseteq X$ of taxa;

3: Use Algorithm 12 to calculate the embedded phylogenetic diversity $PD^{\min}_{\mathsf{T}(\mathcal{N})}(S)$ and $PD^{\max}_{\mathsf{T}(\mathcal{N})}$ for all subsets $S \subseteq X$ of taxa;

4: Use Algorithm 8 to calculate the generalized original Shapley Values $SV_{PND}(a), SV_{PD^{\min}_{\mathsf{T}(\mathcal{N})}}(a)$ and $SV_{PD^{\max}_{\mathsf{T}(\mathcal{N})}}(a)$ for all taxa $a \in X$;

5:

6: **Case 2:** $\mathcal{PD} \in \{PD^{\sum}_{\mathsf{T}(\mathcal{N})}, PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}\}$

7: Use Algorithm 14 to calculate the embedded original Shapley Values $SV^{\sum}_{\mathsf{T}(\mathcal{N})}(a)$ and $SV^{\varnothing}_{\mathsf{T}(\mathcal{N})}(a)$ for all taxa $a \in X$;

8: Set $SV_{PD^{\sum}_{\mathsf{T}(\mathcal{N})}}(a) := SV^{\sum}_{\mathsf{T}(\mathcal{N})}(a)$ and $SV_{PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}}(a) := SV^{\varnothing}_{\mathsf{T}(\mathcal{N})}(a)$;

9:

10: **Case 3:** $\mathcal{PD} = PD^{LSA}$

11: Use Algorithm 15 to calculate the LSA associated original Shapley Value $SV^{LSA}(a)$ and set $SV_{PD^{LSA}}(a) := SV^{LSA}(a)$ for all taxa $a \in X$;

12:

13: **return** $SV_{\mathcal{PD}}(a)$ with $\mathcal{PD} \in \{PND, PD^{\min}_{\mathsf{T}(\mathcal{N})}, PD^{\max}_{\mathsf{T}(\mathcal{N})}, PD^{\sum}_{\mathsf{T}(\mathcal{N})}, PD^{\varnothing}_{\mathsf{T}(\mathcal{N})}, PD^{LSA}\}$ for all $a \in X$;

6.3. Implementation

For the calculation of the various measures of generalized phylogenetic diversity and generalized biodiversity indices we now introduce the software tool `net_diversity.pl`, written in the programming language Perl (v5.18.2), including modules from BioPerl (1.6.923-1) (Vos et al. [34]). The script was tested under the Linux Distribution *Linux Mint 17.2 Rafaela* on a 64-bit computer with an Intel® Core™ i5-430M processor.

We shortly mention some of the modules used for the script, before describing the script itself and analyzing its performance.

Modules

In order to run the script, the following modules are needed:

- `strict` (a Perl pragma to restrict unsafe constructions),
- `warnings` (a Perl pragma to control optional warnings),
- `Try::Tiny` (a module for error handling),
- `Getopt::Long` (a module for the processing of command line options),
- `List::MoreUtils` (a module for functionality on lists),
- `List::Compare` (a module for the comparison of lists),
- `Bio::PhyloNetwork` (a module for phylogenetic networks; Cardona et al. [7]).

The last module is part of the BioPerl package and provides some helpful tools when dealing with phylogenetic networks. Methods used in this project are:

- `$net = Bio::PhyloNetwork->new(-eNewick => $newick_string)` to create a new `Bio::PhyloNetwork` object from its extended Newick representation given in the string `$newick_string`,
- `$net->nodes()` to retrieve the set of all nodes of the network,
- `$net->leaves()` to retrieve the set of all leaves of the network,
- `$net->hybrid_nodes()` to retrieve the set of all reticulation nodes of the network,
- `$net->graph()` to retrieve the underlying graph of the network,
- `$net->explode()` to retrieve the set of all trees (not necessarily phylogenetic X-trees) displayed by the network,

- `$net->eNewick_full()` to retrieve the extended Newick representation of the network as a string.

The module `Bio::PhyloNetwork` relies on the module `Graph::Directed`, because the underlying structure of a phylogenetic network is a directed graph. Irrespective of the methods used by the `Bio::PhyloNetwork` module itself, we additionally use the following methods in this project:

- `$graph->transpose_graph` to reverse the direction of all edges of the graph,
- `$graph->add_edge($u,$v)$` to add the edge (u, v) to the graph,
- `$graph->delete_edge($u,$v)` to delete the edge (u, v) from the graph,
- `$graph->delete_vertex($u)` to delete the node u,
- `$graph->predecessors($u)` to retrieve the parent nodes of a node u,
- `$graph->successors($u)` to retrieve the immediate successors (i.e. the children) of the node u,
- `$graph->topological_sort` to retrieve a linear ordering of the nodes.

Additionally, several methods for dealing with phylogenetic trees are used, which originate in the modules `Bio::TreeIO`, `Bio::Tree::Tree`, `Bio::Tree::TreeFunctionsI` and `Bio::Tree::NodeI`:

- `Bio::TreeIO` is a parser for tree files and creates `Bio::Tree::TreeI` objects (i.e. objects representing phylogenetic trees).
- `Bio::Tree::Tree` allows to access several characteristics of a phylogenetic tree. Methods used in this project are:
  - `$tree->get_nodes()` to retrieve the set of all nodes of the tree,
  - `$tree->get_root_node()` to access the root of the tree,
  - `$tree->get_leaf_nodes()` to retrieve the set of all leaves of the tree,
  - `$tree->as_text('newick')` to retrieve the Newick representation of the tree as a string.
- `Bio::Tree::TreeFunctionsI` provides additional methods for phylogenetic trees, in particular:

  - `$tree->get_lineage_nodes($node)` to retrieve all ancestors of a node (from the root to the most recent ancestor),
  - `$tree->contract_linear_paths` to remove all nodes with in- and outdegree 1 from the tree. Note that this method only works accurately for unweighted phylogenetic trees, i.e. trees without branch lengths. In order to use this functionality for weighted trees, we redefine the method.

- `Bio::Tree::NodeI` provides tools to get information about the nodes of a phylogenetic tree. Methods used in this project are:
  - `$node->get_all_Descendents()` to retrieve all descendants (not just the direct descendants) of a node,
  - `$node->is_Leaf()` to check whether a node is a leaf,
  - `$node->depth()` to retrieve the distance from the root to this node,
  - `$node->branch_length()` to retrieve the edge length between a node and its direct ancestor and `$node->branch_length(value)` to set this edge length to a new value,
  - `$node->id()`, which returns the human readable identifier of a node, e.g. the species name of a leaf.

net_diversity.pl Depending on the options set by the user, `net_diversity.pl` computes one of the several measures of generalized *phylogenetic diversity* for all subsets of taxa of a phylogenetic hybridization network or one of the generalized biodiversity indices for all taxa.

The command

```
~$ ./net_diversity.pl --help
```

yields an overview of the possible options to be chosen when running the script:

```
net_diversity.pl

  Takes a rooted binary phylogenetic hybridization network as input and
  computes several measures of generalized phylogenetic diversity for all
  subsets of taxa and several generalized biodiversity indices for all
  taxa. The network must be represented in the extended Newick format
  (http://dmi.uib.es/~gcardona/BioInfo/enewick.html).

  Please make sure that BioPerl is installed on your machine!

SYNOPSIS:
net_diversity.pl --in=filename --measure=value

OPTIONS:

  --in=filename
    Path to the input file containing the network(s)
    in the extended Newick format.

  --measure=value
    Choose the measure of generalized phylogenetic diversity or
    generalized biodiversity index to calculate.

    Options for measure:
       0 phylogenetic net diversity
       1 embedded phylogenetic diversity (min)
       2 embedded phylogenetic diversity (max)
       3 embedded phylogenetic diversity (sum)
       4 embedded phylogenetic diversity (mean)
       5 LSA associated phylogenetic diversity

       6 embedded Fair Proportion Index (min)
```

```
 7 embedded Fair Proportion Index (max)
 8 embedded Fair Proportion Index (sum)
 9 embedded Fair Proportion Index (mean)
10 LSA associated Fair Proportion Index

11 embedded original Shapley Value (min)
12 embedded original Shapley Value (max)
13 embedded original Shapley Value (sum)
14 embedded original Shapley Value (mean)
15 LSA associated Shapley Value

16 embedded modified Shapley Value (min)
17 embedded modified Shapley Value (max)
18 embedded modified Shapley Value (sum)
19 embedded modified Shapley Value (mean)
20 LSA associated modified Shapley Value

21 embedded unrooted rooted Shapley Value (min)
22 embedded unrooted rooted Shapley Value (max)
23 embedded unrooted rooted Shapley Value (sum)
24 embedded unrooted rooted Shapley Value (mean)
25 LSA associated unrooted rooted Shapley Value

26 generalized original Shapley Value (PND)
27 generalized original Shapley Value (PD_TN^min)
28 generalized original Shapley Value (PD_TN^max)
29 generalized original Shapley Value (PD_TN^sum)
30 generalized original Shapley Value (PD_TN^mean)
31 generalized original Shapely Value (PD^LSA)

32 generalized modified Shapley Value (PND)
33 generalized modified Shapley Value (PD_TN^min)
34 generalized modified Shapley Value (PD_TN^max)
35 generalized modified Shapley Value (PD_TN^sum)
36 generalized modified Shapley Value (PD_TN^mean)
37 generalized modified Shapely Value (PD^LSA)

DESCRIPTION:

  Example: net_diversity.pl --in=myNetwork  --measure=0
```

Performance In order to test the performance of `net_diversity.pl` we randomly generated phylogenetic hybridization networks for different numbers of taxa and reticulation nodes, using the Perl module `Bio::PhyloNetwork::RandomFactory`. This module generates unweighted, binary, tree-child networks. In a second step, we therefore modified the extended Newick representation of the networks and set all branch lengths to 1, respectively. Additionally we used `Bio::PhyloNetwork::TreeFactory` to generate random phylogenetic networks without reticulation nodes, i.e. random phylogenetic trees.

We then analyzed the performance of `net_diversity.pl` for the random networks using the Perl module `Memory::Usage`. However, we did not test all options of `net_diversity.pl`, but chose some representative examples. Additionally, we also analyzed the construction of the (multi)set $\mathsf{T}(\mathcal{N})$ of trees displayed by a network and the construction of the *LSA tree* associated with the network, as these are fundamental underlying methods of the script.

However, the aim of this analysis is not to provide absolute numbers for the computation time and memory usage, but to indicate an overall tendency. Table 9 and Table 10 contain the results of the analysis. The computation time is given in seconds, the memory usage in kilobytes, respectively. Some of the analyses were skipped, indicated by a minus symbol (-) in the corresponding table cells.

Table 9: Performance of `net_diversity.pl` for 10 Taxa

	10 Taxa					
	0 reticulation nodes		5 reticulation nodes		9 reticulation nodes	
Method	Time (sec)	Memory (kb)	Time (sec)	Memory (kb)	Time (sec)	Memory (kb)
PND	< 1	800	4	4024	65	53 986
PD^{LSA}	< 1	2592	< 1	2992	1	3393
$FP^{\varnothing}_{\mathsf{T}(\mathcal{N})}$	< 1	528	< 1	1200	10	13 052
FP^{LSA}	< 1	2600	< 1	2998	< 1	3256
$\widehat{SV}^{\varnothing}_{\mathsf{T}(\mathcal{N})}$	< 1	528	1	1200	10	13 052
$\widehat{SV}^{LSA}$	1 1	2600	< 1	2998	< 1	3256
SV_{PND}	< 1	800	4	4024	65	54 000
$\mathsf{T}(\mathcal{N})$	< 1	524	1	1192	9	12 400
$\mathcal{T}_{LSA}(\mathcal{N})$	< 1	2592	< 1	2992	< 1	3256

Table 10: Performance of `net_diversity.pl` for 20 Taxa

	20 Taxa							
	0 reticulation nodes		5 reticulation nodes		10 reticulation nodes		15 reticulation nodes	
Method	Time (sec)	Memory (kb)	Time (sec)	Memory (kb)	Time (sec)	Memory (kb)	Time (sec)	Memory (kb)
PND	393	335 524	-	-	-	-	-	-
PD^{LSA}	391	156 700	178	186 932	144	186 976	117	186 944
$FP^{\varnothing}_{\mathsf{T}(\mathcal{N})}$	< 1	660	1	2116	32	49 748	1169	1 598 216
FP^{LSA}	< 1	856	< 1	3252	< 1	3784	< 1	5232
$\widehat{SV}^{\varnothing}_{\mathsf{T}(\mathcal{N})}$	< 1	664	1	2116	33	49 760	1238	1 601 660
$\widehat{SV}^{LSA}$	< 1	2856	< 1	3252	< 1	3780	< 1	5232
SV_{PND}	477	402 056	-	-	-	-	-	-
$\mathsf{T}(\mathcal{N})$	< 1	660	1	1980	28	47 632	1063	1 533 940
$\mathcal{T}_{LSA}(\mathcal{N})$	< 1	2856	< 1	3256	< 1	3780	< 1	4700

Remarks.

- Comparing the computation of the (multi)set $\mathsf{T}(\mathcal{N})$ of trees displayed by a network $\mathcal{N}$ and the construction of the *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N})$, we see that the latter requires more memory for networks with up to five reticulation nodes. However, when the number of reticulation nodes grows, the memory usage for the computation of the (multi)set $\mathsf{T}(\mathcal{N})$ rises exponentially, while it only marginally increases for the *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N})$.
  The same holds for the computation time: While the computation time for the (multi)set $\mathsf{T}(\mathcal{N})$ increases exponentially, the computation time for the *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N})$ remains low.
  As the random networks used in this analysis were binary, tree-child networks, the results for the (multi)set $\mathsf{T}(\mathcal{N})$ come up to our expectations, because there are $2^{r(\mathcal{N})}$ trees displayed by a tree-child networks with $r(\mathcal{N})$ reticulation nodes.
  For the construction of the *LSA tree* $\mathcal{T}_{LSA}(\mathcal{N})$ additional analyses for a hybridization network with 50 taxa and 49 reticulation nodes, as well as for a hybridization network with 75 taxa and 74 reticulation nodes and a hybridization network with 100 leaves and 50 reticulation nodes were conducted, resulting in computation times of 3, 9 and 4 seconds, respectively. We suspect the computation times for the construction of the *LSA tree* to stay low for even bigger networks with more taxa and more reticulation nodes. However, due to a high computation time of the

random network generation for greater numbers of taxa and reticulation nodes, this could not be tested.
Additionally, the Perl module `Bio::PhyloNetwork::RandomFactory` only allows for the generation of tree-child network. Thus, it remains to analyze the computation times for the *LSA tree* for more general, non tree-child, networks.

- If we compare the computation times for the *phylogenetic net diversity* PND and the *generalized Shapley Value* SV_{PND}, we see that they are almost identical or at least of the same dimension. This suggests that the complexity in calculating the *Shapley Value* according to its definition primarily arises from the need of computing the *phylogenetic diversity* for all subsets of taxa.

- Similar results hold for the computation times of generalized biodiversity indices based on the (multi)set $\mathsf{T}(\mathcal{N})$ of embedded trees and the computation time for $\mathsf{T}(\mathcal{N})$ itself. Comparing $FP^{\varnothing}_{\mathsf{T}(\mathcal{N})}$, $\widehat{SV}^{\varnothing}_{\mathsf{T}(\mathcal{N})}$ and $\mathsf{T}(\mathcal{N})$ for a fixed number of taxa and a fixed number of reticulation nodes, we see that their computation times are of the same dimension. As the number of embedded trees grows exponentially, so do the computation times for $FP^{\varnothing}_{\mathsf{T}(\mathcal{N})}$ and $\widehat{SV}^{\varnothing}_{\mathsf{T}(\mathcal{N})}$.

- Comparing the computation times for the calculation of PND for a phylogenetic tree with 10 taxa and a phylogenetic tree with 20 taxa (i.e. for the two networks without reticulation nodes), we see an increase from < 1 to 393 seconds. An additional analysis of a tree with 21 taxa resulted in a computation time of 867 seconds. Thus, we see an exponential growth in the computation times. This is due to the fact that `net_diversity.pl` calculates the PND of all subsets $S \subseteq X$, which are $2^{|X|}$ many.
  Thus, the computation of PND may already be time-consuming for a single phylogenetic X-tree. In case of phylogenetic networks with at least one reticulation node, it has to be repeated $|\overline{\mathsf{T}(\mathcal{N})}|$ times. This is illustrated by the rise in computation times for PND for the network with 10 taxa and 0, 5 and 9 reticulation nodes from < 1, over 4 to 65 seconds, respectively. The same is expected for the network with 20 taxa, but the analysis is skipped.
  In contrast to the computation of PND, the computation of PD^{LSA} has to be conducted only once, namely for the *LSA tree*. If the network under consideration has no reticulation nodes, i.e. it is already a tree, the computation times for PND and PD^{LSA} should be almost identical, which is the case for both networks.
  Surprisingly at first glance, the computation time for PD^{LSA} decreases with an increasing number of reticulation nodes for the second network, even though the number of taxa equals 20 in all cases. This can, however, be explained by the

structure of the *LSA tree*. With an increasing number of reticulation nodes, the *LSA tree* tends to become more and more unresolved (cf. Figure 30). This seems to facilitate the computation of PD for all subsets of taxa according to Algorithm 4.

Summarizing the above, we see that there are two major causes for high computation times of `net_diversity.pl`. On the one hand, an increasing number of taxa causes an increase in the number of subsets of taxa to be considered when calculating any measure of generalized *phylogenetic diversity*. For a set X, there are $2^{|X|}$ subsets $S \subseteq X$ and thus, the number of subsets grows exponentially with an increasing size $|X|$ of the taxon set. This makes the calculation of *phylogenetic diversity* and biodiversity indices explicitly based on the *phylogenetic diversity* of subsets of taxa infeasible even for large phylogenetic X-trees (cf. Wicke [36]). For phylogenetic networks, however, the calculations become even 'more infeasible', if the *phylogenetic diversity* of all subsets of taxa has to be computed for all trees in $\mathsf{T}(\mathcal{N})$ (in case of the *embedded phylogenetic diversity* and biodiversity indices based on this measure) or in $\overline{\mathsf{T}(\mathcal{N})}$ (in case of the *phylogenetic net diversity* and biodiversity indices based on this measure).

Additionally, computation times for all measures of *embedded phylogenetic diversity* or *embedded biodiversity indices* are strongly influenced by the number of reticulation nodes of a network. In order to retrieve the (multi)set $\mathsf{T}(\mathcal{N})$ of phylogenetic X-trees displayed by a network, the (multi)set $\overline{\mathsf{T}(\mathcal{N})}$ of all trees embedded in $\mathcal{N}$ has to be computed, before all trees that are not phylogenetic X-trees can be discarded. However, for a binary phylogenetic network $\mathcal{N}$ on X with $r(\mathcal{N})$ reticulation nodes, there are $2^{r(\mathcal{N})}$ embedded trees, i.e. $|\overline{\mathsf{T}(\mathcal{N})}| = 2^{r(\mathcal{N})}$. Thus, the size of the (multi)set $\overline{\mathsf{T}(\mathcal{N})}$ grows exponentially with an increasing number of reticulation nodes. This makes all calculations based on $\overline{\mathsf{T}(\mathcal{N})}$ or $\mathsf{T}(\mathcal{N})$ infeasible for phylogenetic networks with a high number of reticulation nodes.

The only methods that seem to be feasible for large phylogenetic networks and high numbers of reticulation nodes are biodiversity indices based on the *LSA tree*. However, the *LSA tree* associated with a network reduces the network to its most basic treelike content and disregards much of the structure of the network. In this case, fast computation times come at the expense of loss of structure of the network and thus, of loss of evolutionary information.

However, in the following we use `net_diversity.pl` to calculate different generalized biodiversity indices for the taxa of a hybridization network for oceanic dolphins.

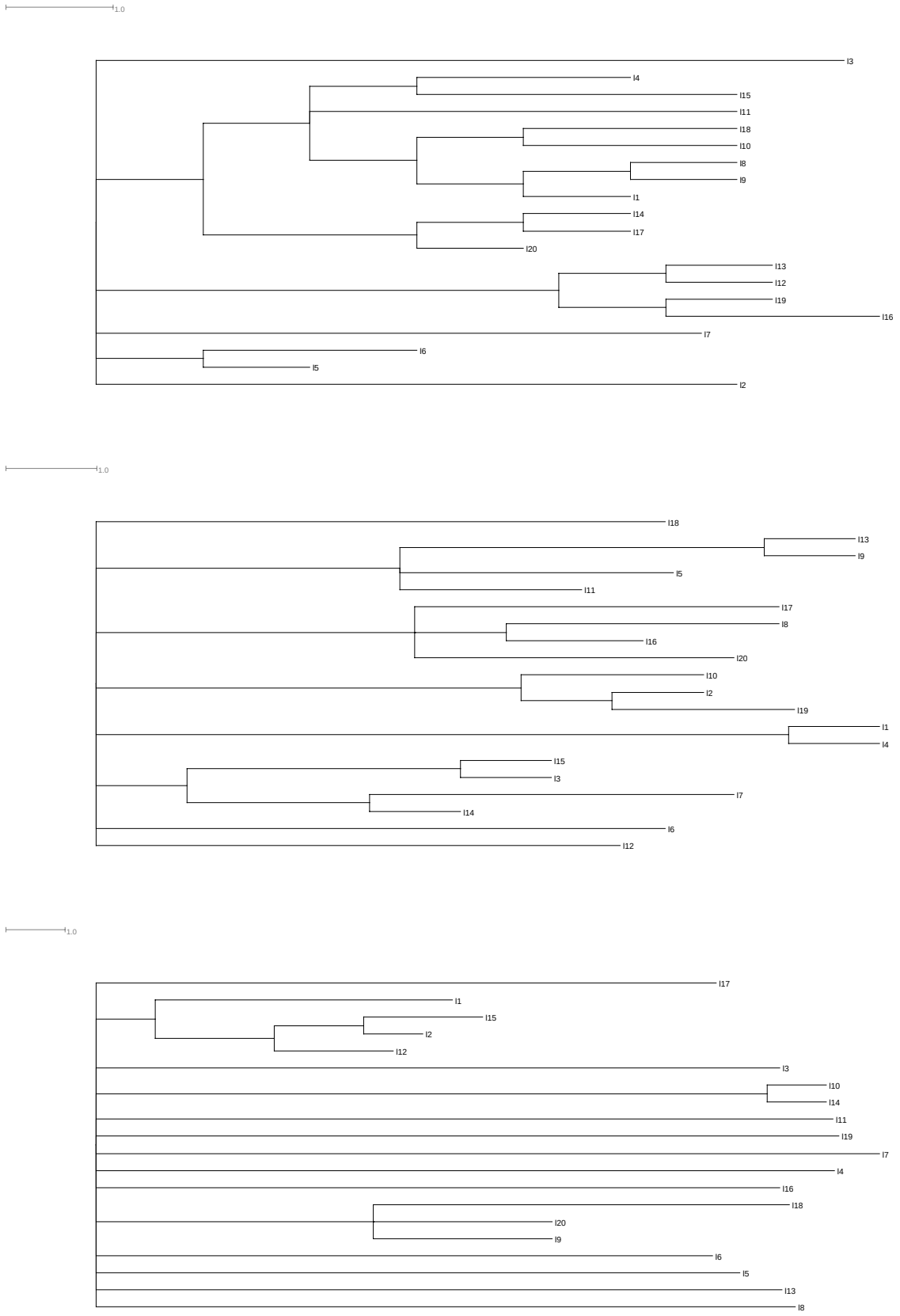


Fig. 30: *LSA trees* for random binary hybridization networks with 20 taxa and 5 (top), 10 (middle), 15 (bottom) reticulation nodes (created with Dendroscope: Huson and Scornavacca [19]).

7. Example – a dolphin data set

Hybridization has been mainly linked to plant biology in the past. However, in more recent times hybridization has been documented or hypothesized not only to occur in plants, but also in animals, e.g. in insects (Fontaine et al. [14], Wen et al. [35]), gastropods (Haase et al. [17], Zielske and Haase [38]) and sharks (Marino et al. [26], Morgan et al. [28]).

In the following we consider hybrid speciation in dolphins, i.e. in marine mammals, based on a study by Amaral et al. [5]. We will shortly outline the contents of this study and describe our own approach of inferring a hybridization network for the taxa of this study. We will then use this hybridization network as an example for the calculation of generalized biodiversity indices. However, as we could not directly construct a hybridization network from the data of this study, but inferred it manually, we emphasize that this example only serves as a vivid illustration for the concepts of generalized *phylogenetic diversity* and generalized biodiversity indices, while, biologically, it requires further and more thorough examination.

7.1. Hybrid speciation in dolphins

In their study, Amaral et al. [5] consider a group of marine mammals, namely the family Delphinidae. Delphinidae belong to the order of Cetaceans, whose evolution has been characterized by 'some rapid radiation events, which in some groups has led to confusing taxonomy and a difficulty in clarifying phylogenetic relationships due to the confounding effects of incomplete lineage sorting and possibly hybridization' (Amaral et al. [5]). The authors turn their attention to the latter and assume a hybrid origin for the Clymene Dolphin (*Stenella clymene* (Gray, 1850)[14]).

Stenella clymene is endemic to the Atlantic Ocean and while its 'cranial features closely resemble those of *Stenella coeruleoalba*, [...] its external appearance and behavior are more similar to those of *Stenella longirostris*' (Amaral et al. [5]). Moreover, Amaral et al. [5] observe a strong discordance between mitochondrial and nuclear markers, the former suggesting that *S. clymene* is more closely related to *S. coeruleoalba* (Meyen, 1833), while the latter suggests a closer relationship of *S. clymene* to *S. longirostris* (Gray, 1828). The authors admit that this discordance between different markers could also be due to processes such as incomplete lineage sorting, but according to Amaral et al. [5] it does not explain it entirely. They come to the conclusion that the 'discrepancy between mitochondrial markers and nuclear markers suggests a

[14] Taxonomic information derived from the *Integrated Taxonomic Information System* (`https://www.itis.gov/`).

hybrid origin of *S. clymene*, as a result of an ancient hybridization between a female *S. coeruleoalba* and a male *S. longirostris*' (Amaral et al. [5]).

7.2. Inference of a phylogenetic network

For the mitochondrial DNA Amaral et al. [5] present a Bayesian phylogenetic tree based on the cytochrome *b* gene (cf. Figure 31). The tree contains several individuals of the species *S. clymene*, *S. longirostris* and *S. coerluleoalba*. Moreover, it contains samples for *Delphinus delphis* (Linneaus, 1758), *Delphinus capensis* (Gray, 1828), *Delphinus capensis tropicalis* (van Bree, 1971), *Stenella attenuata* (Gray, 1846), *Stenella frontalis* (Cuvier, 1829), *Tursiops truncatus* (Montagu, 1821), *Tursiops aduncus* (Ehrenberg, 1833), *Lagenodelphis hosei* (Fraser, 1956), *Sousa chinensis* (Osbeck, 1756), *Sotalia fluviatilis* (Gervais and Deville, 1853), *Globicephala melas* (Trail, 1809) and *Phocoena phocoena* (Linneaus, 1758)[15].

Additionally, a Bayesian species tree based on nine nuclear loci (Del_04, Del_05, Del_10, Del_11, Del_12, Del_16, BTN, PLP, CHRNA1; cf. Amaral et al. [4]) is presented (cf. Figure 32).

While the tree based on nuclear DNA places *S. clymene* and *S. longirostris* closer together, the tree based on mitochondrial DNA shows that most of the *S. clymene* samples are closely related to *S. coeruleoalba*. There are, however, some *S. clymene* samples placed near *S. longirostris*. Note that the species tree in Figure 32 is an ultrametric tree[16] with branch lengths representing time (Amaral et al. [5] have enforced a strict molecular clock), while branch lengths in the mitochondrial tree represent substitutions per site and the tree is not ultrametric (cf. Figure 31).

We therefore inferred another ultrametric species tree, solely based on the cytochrome *b* gene, i.e. on mitochondrial DNA. In order to do so, we used data deposited in the Dryad Repository (`http://dx.doi.org/10.5061/dryad.6dr0475t`) by Amaral et al. [4]. Amongst others, this data set contained sequences of the cytochrome *b* gene for all species listed above, except for *Stenella clymene*. In some cases there were two samples for one species (e.g. *Delphinus capensis* I and *Delphinus capensis* II), so we first reduced the data set to one sample per species. We then retrieved a sequence of the cytochrome *b* gene for *Stenella clymene* from GenBank (GenBank Accession Number AF084083.1) and added it to the data set, which now comprised 15 sequences. The sequence for

[15] *Phocoena phocoena* does not belong to the family Delphinidae, but to the family Phocoenidae and is used as an outgroup here.

[16] A rooted phylogenetic tree is called ultrametric if the path lengths from the root to each leaf are identical.

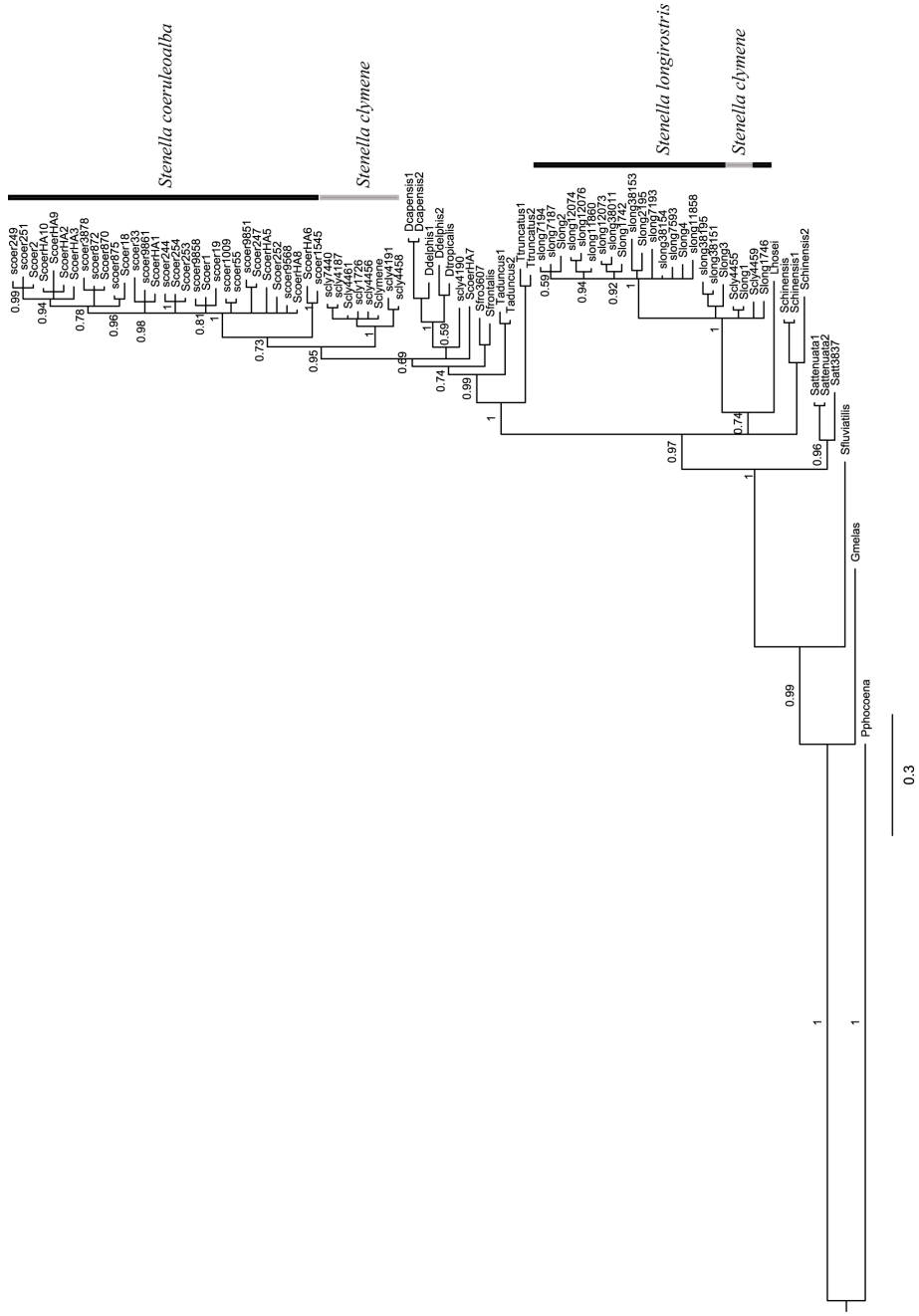


Fig. 31: Bayesian phylogenetic tree generated in MrBayes for the cytochrome *b* gene. Posterior probability values are above nodes. Branch lengths are in substitutions/site (Figure taken from Amaral et al. [5]).

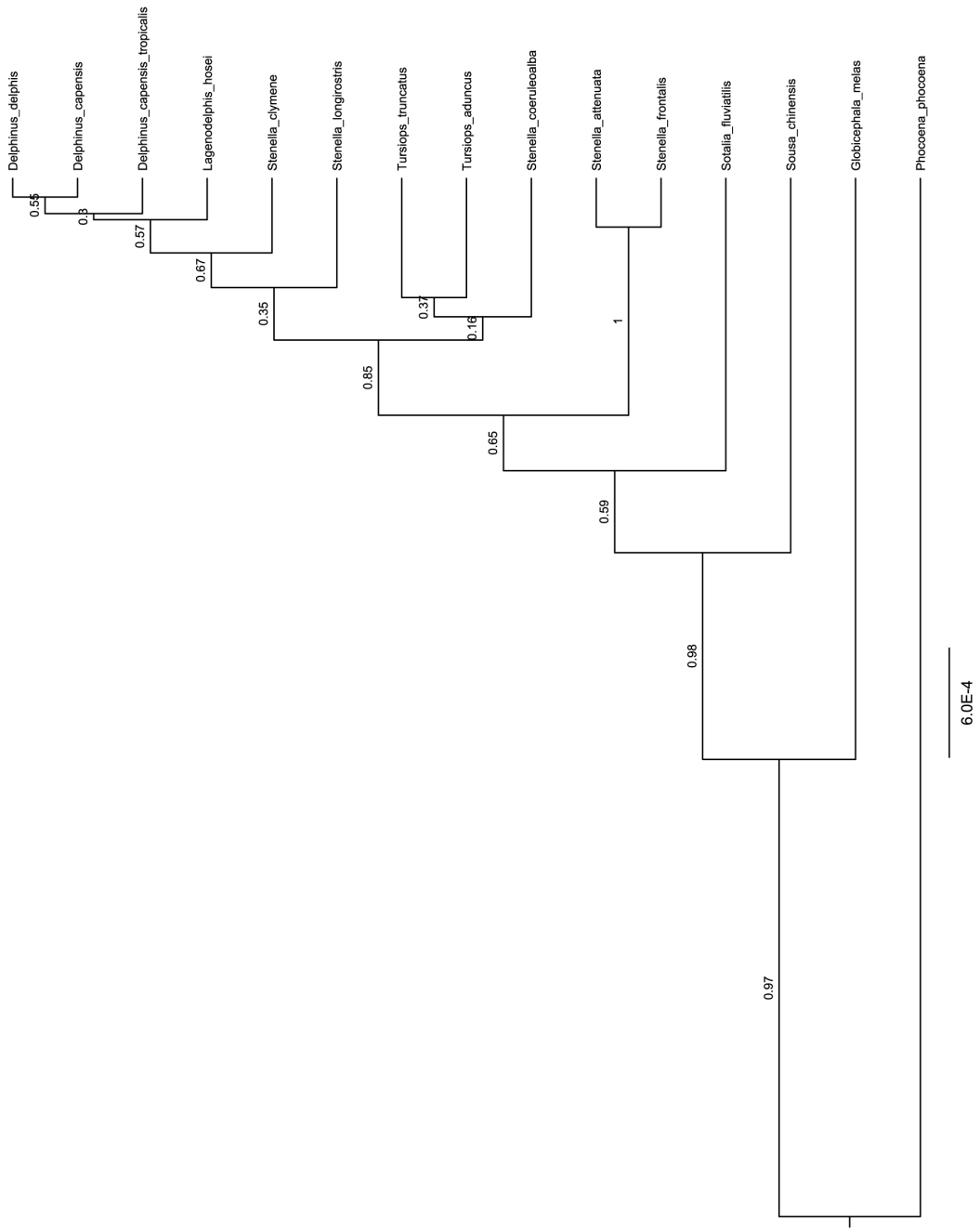


Fig. 32: Species tree estimated with the *BEAST method. Posterior probability values are above nodes (Figure taken from Amaral et al. [5]).

Stenella Clymene was 21 nucleotides longer than that of the other species in the data set, so we produced a ClustalW alignment in BioEdit v.7.2.5 with the default settings. As a result 21 gaps had to be introduced at the beginning of all sequences (except for the *S. clymene* sequence), while the rest of the sequences was aligned without gaps. The data set was then used to estimate a Bayesian phylogenetic tree using the program MrBayes v.3.2 (Ronquist et al. [29]). 1 million MCMC generations, sampling every 100 generations were run. Following Amaral et al. [5], the data set was partitioned by codon positions and the Tamura-Nei model was chosen as nucleotide substitution model.[17] The sequence from *Phocoena phocoena* was used as outgroup and a strict molecular clock was enforced.

The resulting tree is depicted in Figure 33. As suggested by Amaral et al. [5], the mitochondrial DNA of *S. clymene* is more closely related to *S. coeruleoalba* than to *S. longirostris*. Our tree resembles the tree for the mitochondrial DNA from Amaral et al. [5] (cf. Figure 31), even though there are small differences, which may deserve further revision.

However, we then used Dendroscope v.3.5.7 (Huson and Scornavacca [19]) to calculate a minimum hybridization network from the species tree based on nuclear data (Figure 32) and the species tree based on mitochondrial data (Figure 33). This calculation resulted in a total of 25 phylogenetic networks with a reticulation number of 6, respectively. 22 out of 25 networks suggested a hybrid origin for *Stenella clymene*, but *S. longirostris* was never among its putative parents, while *S. coeruleoalba* was in most cases. On the contrary, in 20 times the networks suggested a hybrid origin for *S. longirostris* as well, in 18 times thereof simultaneously to a hybrid origin for *S. clymene*.

These results, in particular the high number of reticulations in the networks, can be explained by the fact that the species tree based on nuclear DNA and the species tree based on mitochondrial DNA show more discordance than the placement of *Stenella clymene* (cf. Figures 32 and 33). We suppose that this is not only due to possible hybridization events, but also to processes such as incomplete lineage sorting. Thus, in order to obtain realistic networks, both incomplete lineage sorting and hybridization should be taken into account. Moreover, the species tree obtained from the cytochrome *b* gene could not be fully resolved, which may be another confounding aspect. Last but not least, Dendroscope sets all edge lengths to 1 in the process of calculating the hybridization network, which makes it difficult to use the results for our aim of calculating realistic measures of generalized *phylogenetic diversity* and generalized

[17] A analysis of our data set with jModeltest v.2.1.10 (Darriba et al. [10]) also suggested the Tamura-Nei model.

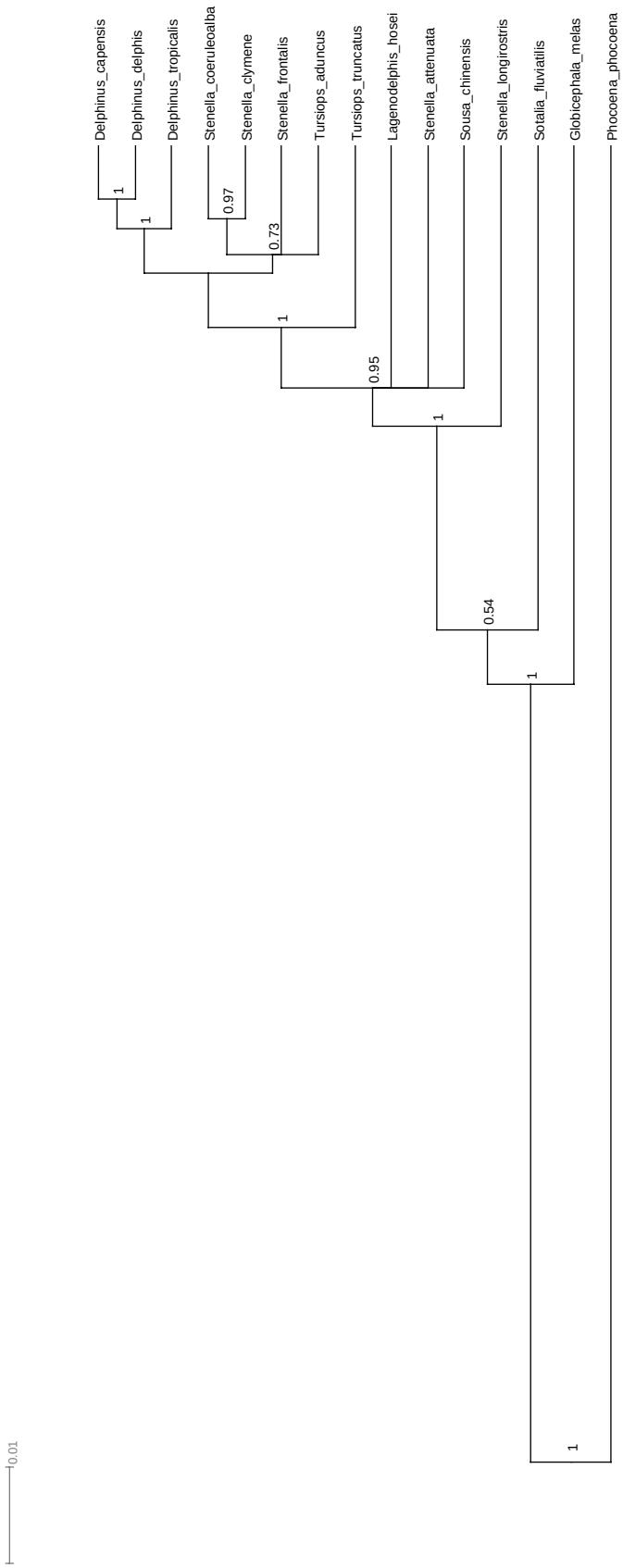


Fig. 33: Species tree (50 % majority rule consensus) for the cytochrome *b* gene estimated with MrBayes. Posterior probabilities are next to nodes.

biodiversity indices.

We therefore decided to manually edit the Delphinidae species tree presented in Amaral et al. [5] and replaced the edge directed into *S. clymene* by two new edges, i.e. reticulation edges, between *S. clymene* and *S. coeruleoalba* and *S. longirostris*, respectively (cf. Figure 34). We placed the hybridization event of *S. coeruleoalba* and *S. longirostris* near to the tips of the tree, resulting in a short pending edge for *S. clymene*, as we did not have any information about the time of the hypothesized hybridization event. Amaral et al. [5], however, indicate that '*S. clymene* is currently distinct from its parental species, although backcross may still occur' (Amaral et al. [5]), so we supposed that *S. clymene* was a relatively young species. However, these assumptions are very speculative and we emphasize again that the following calculations of generalized biodiversity indices for the resulting network only serve as an example for the illustration of the concepts introduced in previous chapters, but should not be considered as such in future prioritization decisions concerning the family Delphinidae.

7.3. Calculation of the original Shapley Value

Recall that the different versions of the *Shapley Value (original, modified, unrooted rooted)* and the *Fair Proportion Index* are closely related. In the following we will not further dwell on their relationship, but focus on the impact of considering a phylogenetic hybridization network as opposed to a phylogenetic tree. Thus, in the following we will focus on the *original Shapley Value* and calculate this value for the family Delphinidae. On the one hand, we will base the calculation on both the nucelar species tree (cf. Figure 32) and the mitochondrial species tree (cf. Figure 33). On the other hand, we will calculate it based on the hybridization network depicted in Figure 34. Note, however, that the branch lengths in the mitochondrial species tree are significantly larger than in the nuclear species tree or the hybridization network. Thus, the absolute values for the *original Shapley Value* are not directly comparable. We can, however, compare the ranking order of taxa induced by the *original Shapley Value* for both trees and the network. In case of the phylogenetic network, we will also compare the ranking order induced by the different generalized versions of the *original Shapley Value*.

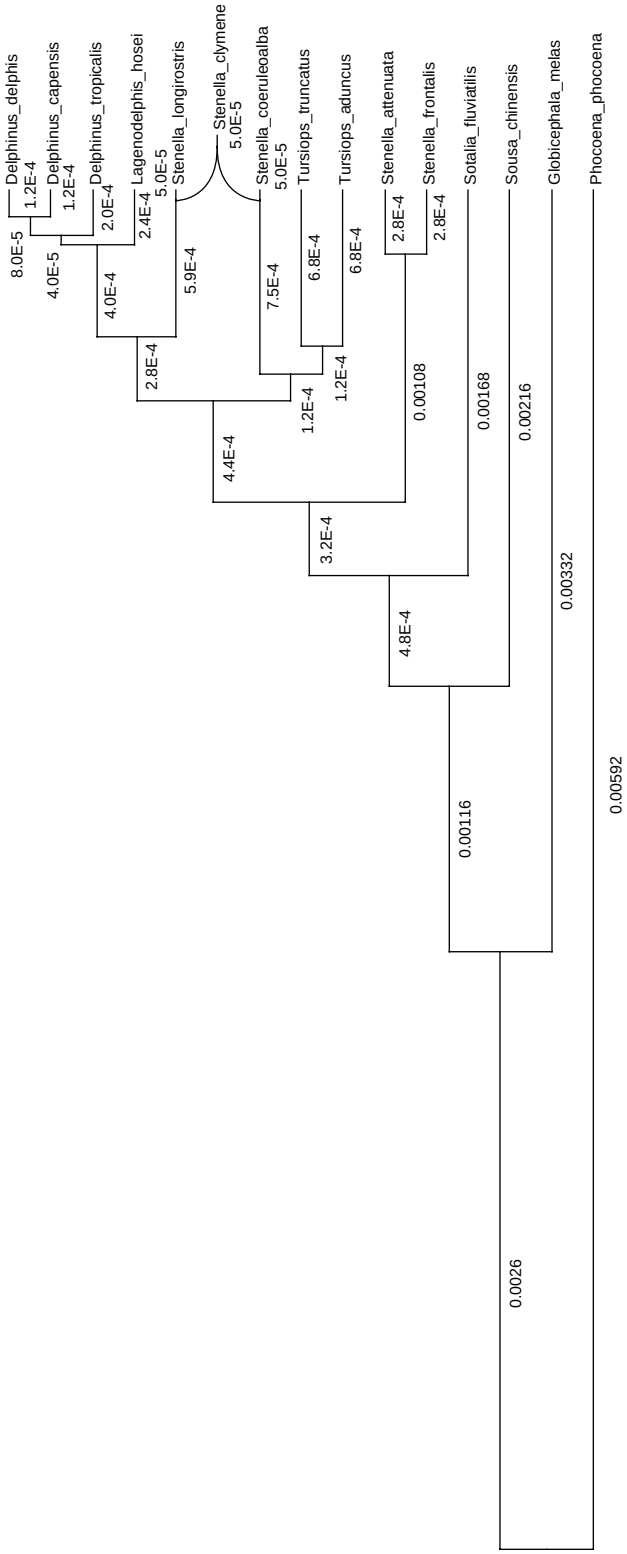


Fig. 34: Hybridization network for the Delphinidae (created with Dendroscope: Huson and Scornavacca [19]).

Original Shapley Value based on the nuclear and mitochondrial species trees

- Original Shapley Values (rounded) based on the nuclear species tree (Figure 32):

Phocoena phocoena	0.005 920
Globicephala melas	0.003 506
Sousa chinensis	0.002 435
Sotalia fluviatilis	0.001 995
Stenella coeruleoalba	0.001 233
Tursiops truncatus	0.001 173
Tursiops aduncus	0.001 173
Stenella frontalis	0.001 164
Stenella attenuata	0.001 164
Stenella longirostris	0.001 080
Stenella clymene	0.000 920
Lagenodelphis hosei	0.000 770
Delphinus tropicalis	0.000 743
Delphinus delphis	0.000 703
Delphinus capensis	0.000 703

- Original Shapley Values (rounded) based on the mitochondrial species tree (Figure 33):

Phocoena phocoena	0.137 287
Globicephala melas	0.061 916
Sotalia fluviatilis	0.056 716
Stenella longirostris	0.037 243
Stenella attenuata	0.033 628
Lagenodelphis hosei	0.033 628
Sousa chinensis	0.033 628
Tursiops truncatus	0.028 143
Stenella frontalis	0.021 811
Tursiops aduncus	0.021 811
Delphinus tropicalis	0.020 133
Stenella clymene	0.019 916
Stenella coeruleoalba	0.019 916
Delphinus capensis	0.018 585
Delphinus delphis	0.018 585

When comparing the rankings induced by the *original Shapley Value* for the species tree based on nuclear DNA and the species tree based on mitochondrial DNA, we see

that only the upper most two positions and the lowest two positions are occupied by the same species (*Phocoena phocoena* and *Globicephala melas* and *Delphinus capensis* and *Delphinus delphis*, respectively). All other taxa swap their positions, due to the discrepancies between the topology of the nuclear species tree and the mitochondrial species tree.

Original Shapley Value based on the hybridization network We will now consider the *original Shapley Value* based on the hybridization network depicted in Figure 34. As indicated above, this network was derived from the nuclear species tree by manually introducing hybridization edges. Note however, that neither the nuclear species tree nor the mitochondrial species tree are displayed by the network. The set of embedded trees for the dolphin network can be found in Figure 36, while the *LSA tree* associated with it, is depicted in Figure 35.

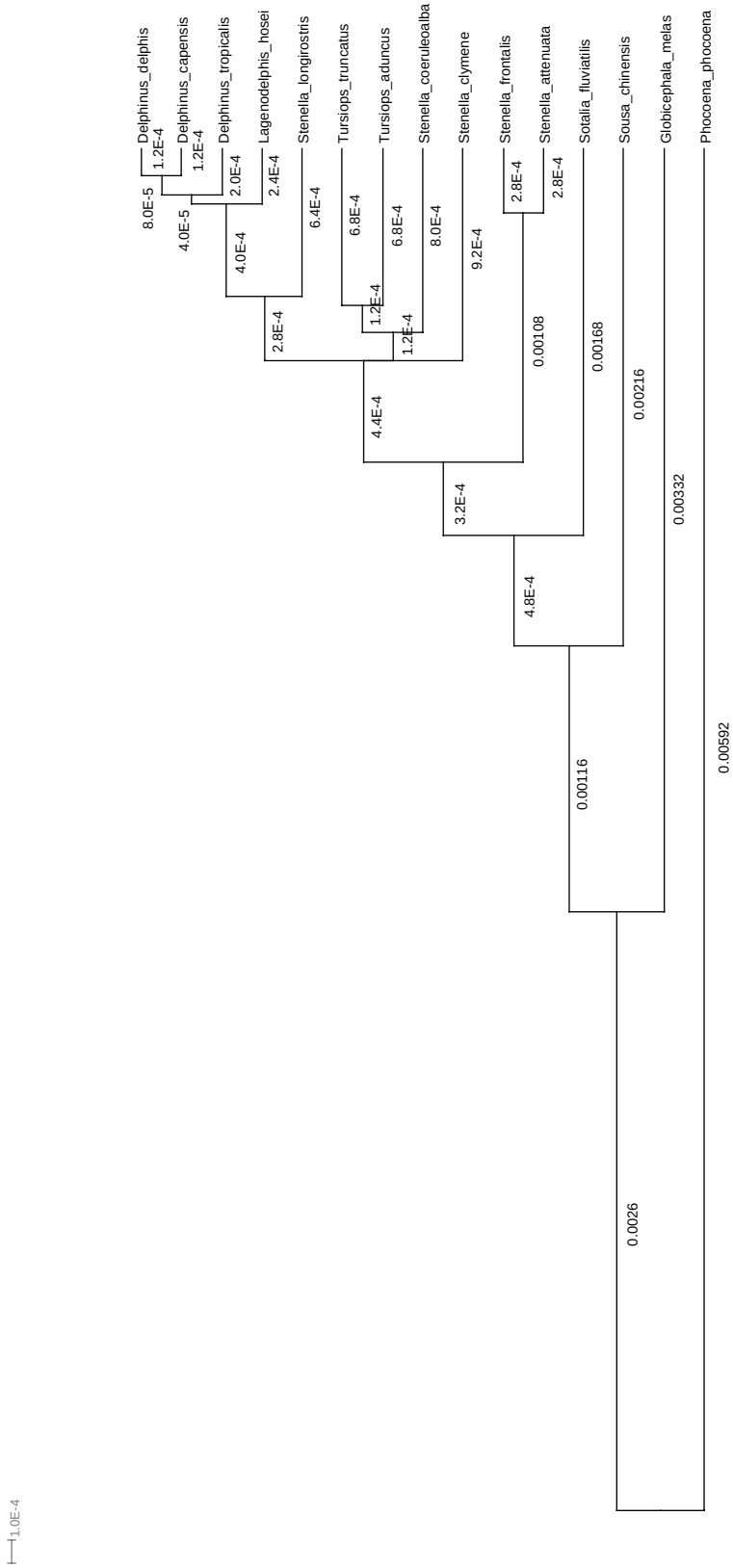


Fig. 35: *LSA tree* associated with the dolphin network (created with Dendroscope: Huson and Scornavacca [19]).

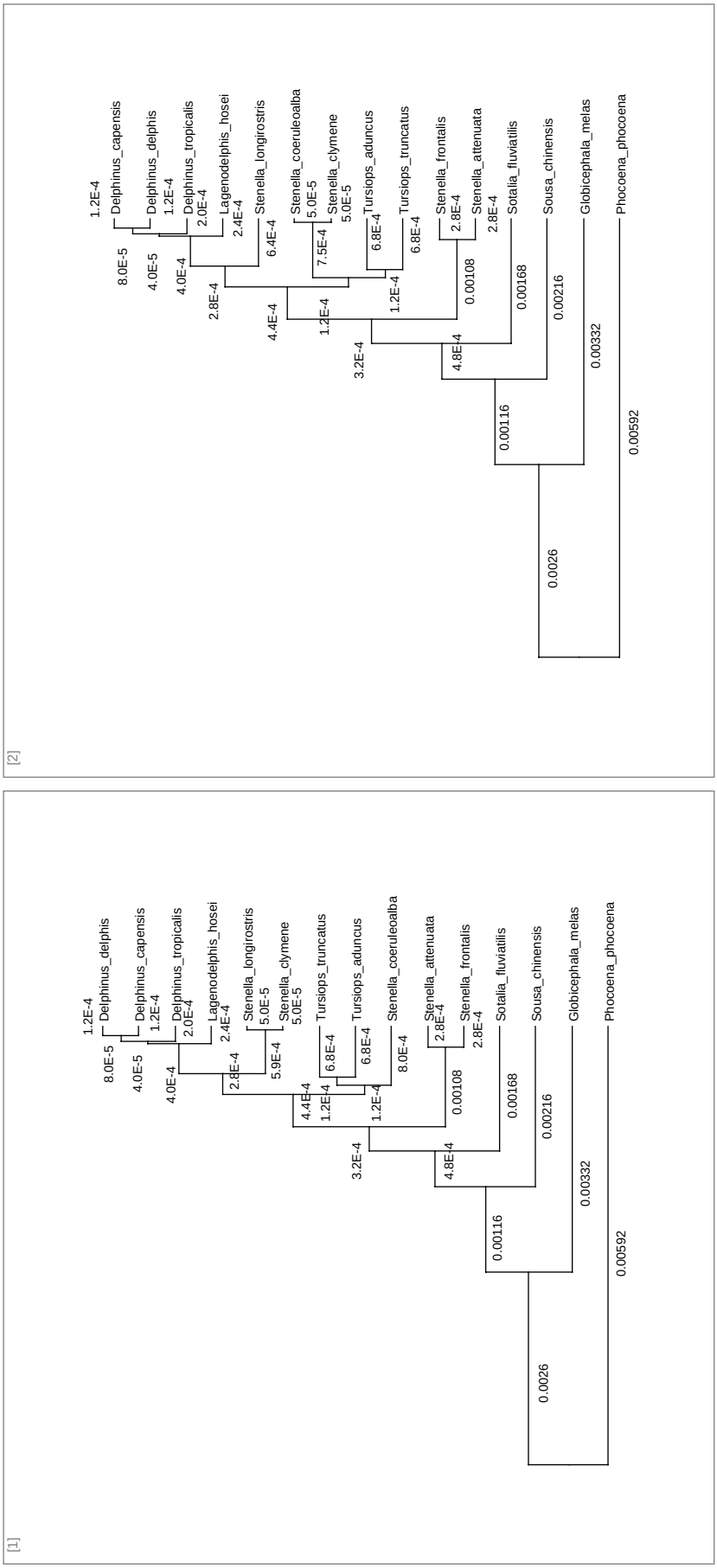


Fig. 36: Trees displayed by the dolphin network (created with Dendroscope: Huson and Scornavacca [19]).

Table 11: Original Shapley Values (rounded) for the dolphin hybridization network. Numbers in parentheses reflect the ranking position of the species.

	$SV^{\min}_{\mathsf{T}(\mathcal{N})}$	$SV^{\max}_{\mathsf{T}(\mathcal{N})}$	$SV^{\sum}_{\mathsf{T}(\mathcal{N})}$	$SV^{\varnothing}_{\mathsf{T}(\mathcal{N})}$	SV^{LSA}	$SV_{PD^{\min}_{\mathsf{T}(\mathcal{N})}}$	$SV_{PD^{\max}_{\mathsf{T}(\mathcal{N})}}$
Phocoena phocoena	0.005 920 (1)	0.005 920 (1)	0.011 840 (1)	0.005 920 (1)	0.005 920 (1)	0.005 920 (1)	0.005 920 (1)
Globicephala melas	0.003 506 (2)	0.003 506 (2)	0.007 011 (2)	0.003 506 (2)	0.003 506 (2)	0.003 506 (2)	0.003 506 (2)
Sousa chinensis	0.002 435 (3)	0.002 435 (3)	0.004 870 (3)	0.002 435 (3)	0.002 435 (3)	0.002 435 (3)	0.002 435 (3)
Sotalia fluviatilis	0.001 995 (4)	0.001 995 (4)	0.003 990 (4)	0.001 995 (4)	0.001 995 (4)	0.001 995 (4)	0.001 995 (4)
Tursiops aduncus	0.001 163 (7)	0.001 173 (6)	0.002 336 (5)	0.001 168 (5)	0.001 173 (7)	0.001 171 (5)	0.001 165 (5)
Tursiops truncatus	0.001 163 (7)	0.001 173 (6)	0.002 336 (5)	0.001 168 (5)	0.001 173 (7)	0.001 171 (5)	0.001 165 (5)
Stenella frontalis	0.001 164 (5)	0.001 164 (8)	0.002 328 (7)	0.001 164 (7)	0.001 164 (9)	0.001 164 (7)	0.001 164 (7)
Stenella attenuata	0.001 164 (5)	0.001 164 (8)	0.002 328 (7)	0.001 164 (7)	0.001 164 (9)	0.001 164 (7)	0.001 164 (7)
Stenella clymene	0.000 785 (10)	0.000 848 (11)	0.001 633 (11)	0.000 816 (11)	0.001 313 (5)	0.000 676 (15)	0.000 957 (9)
Stenella coeruleoalba	0.000 848 (9)	0.001 233 (5)	0.002 081 (9)	0.001 040 (9)	0.001 233 (6)	0.001 129 (9)	0.000 952 (10)
Stenella longirostris	0.000 785 (10)	0.001 089 (10)	0.001 874 (10)	0.000 937 (10)	0.001 089 (11)	0.000 985 (10)	0.000 888 (11)
Lagenodelphis hosei	0.000 780 (12)	0.000 789 (12)	0.001 569 (12)	0.000 784 (12)	0.000 789 (12)	0.000 783 (11)	0.000 785 (12)
Delphinus tropicalis	0.000 753 (13)	0.000 762 (13)	0.001 515 (13)	0.000 758 (13)	0.000 762 (13)	0.000 757 (12)	0.000 758 (13)
Delphinus delphis	0.000 713 (14)	0.000 722 (14)	0.001 435 (14)	0.000 718 (14)	0.000 722 (14)	0.000 717 (13)	0.000 718 (14)
Delphinus capensis	0.000 713 (14)	0.000 722 (14)	0.001 435 (14)	0.000 718 (14)	0.000 722 (14)	0.000 717 (13)	0.000 718 (14)

Results (cf. Table 11):

- The ranking obtained for the network by $SV^{\max}_{\mathsf{T}(\mathcal{N})}$ equals the ranking obtained from SV for the nuclear tree, even though the nuclear tree is not displayed by the network. Thus, this might be a coincidence.

- In accordance with the nuclear species tree, the species *Phocoena phocoena*, *Globicephala melas*, *Sousa chinensis* and *Sotalia fluviatilis* occupy the first four positions in all rankings induced by the *original Shapley Value* for the hybridization network. Moreover, they receive the same absolute value for each version of the *Shapley Value* (except for $SV^{\Sigma}_{\mathsf{T}(\mathcal{N})}$, which is two times this value). This can be explained by the fact that these species are only distantly related to the hybrid species *Stenella clymene* and are not affected by rearrangement of the tree topology (i.e. they are placed at the same position in both embedded trees and the *LSA tree* associated with the network).
  Surprisingly, the two species *Stenella attenuata* and *Stenella frontalis* receive different ranking positions by the different versions of the *Shapley Value*, even though, the absolute values are identical for each version (except for $SV^{\Sigma}_{\mathsf{T}(\mathcal{N})}$) and the two species do not change their position in the set of embedded trees or the *LSA tree*. However, their pending edges are relatively short and they are a sister group to the taxa in the subtree rooted at the lowest common ancestor of *Delphinus capensis* and *Tursiops truncatus*, a group of species that undergoes rearrangement in the set of embedded trees and the *LSA tree*. Thus, the ranking position of *Stenella frontalis* and *Stenella attenuata* is affected by the ranking positions of the taxa in this subtree.

- Again, in accordance with the nuclear species tree, the species *Lagenodelphis hosei*, *Delphinus tropicalis*, *Delphinus delphis* and *Delphinus capensis* are ranked the lowest in most cases, probably due to their short pending edges and their general position in the network, which suggests that this group of species only contributes marginally to overall *phylogenetic diversity*. Except for *Lagenodelphis hosei* this general tendency is also reflected by the ranking induced by the *original Shapley Value* on the mitochondrial species tree.
  Lagenodelphis hosei, however, receives a ranking position further to the front in the ranking obtained from the mitochondrial tree.

- When considering the hybrid species *Stenella clymene* and its parents *Stenella longirostris* and *Stenella coeruleoalba*, we see that almost all versions of the *original Shapley Value* for networks induce a ranking, where these three species

take a position in the middle field. However, for *Stenella clymene* there are two exceptions: SV^{LSA} places it on position five, while $SV_{PD^{\min}_{\mathsf{T}(\mathcal{N})}}$ ranks it the lowest (position fifteen). For comparison, in the nuclear species tree *Stenella clymene* is ranked as number eleven, in the mitochondrial species tree as number twelve.
For *Stenella coeruleoalba* there are also two indices which rank it higher than on average (position five in case of $SV^{\max}_{\mathsf{T}(\mathcal{N})}$ and position six in case of SV^{LSA}), but in contrast to *Stenella clymene* these two exceptions have the same direction (i.e. they both rank *Stenella coeruleoalba* higher than the other versions of the *original Shapley Value* do on average). *Stenella longirostris*, however, receives position ten or eleven by all versions of the *original Shapley Value* for this network, which reflects the situation in the nuclear species tree, but not the one in the mitochondrial species tree, where it is placed on position four.

- Comparing the rankings induced by the different versions of the *original Shapley Value* for the dolphin network, we see that they roughly induce the same ranking order. There are, however, small differences and especially the hybrid species *Stenella clymene* is assessed differently by SV^{LSA} and $SV_{PD^{\min}_{\mathsf{T}(\mathcal{N})}}$ compared to the other indices.
  These results, however, suggest further analysis of the different generalized biodiversity indices and evaluation of their correlation or further examination of the relationship of the rankings they induce.

Even though this example cannot be used in real taxon prioritization decisions (due to the nonscientific way the hybridization network was inferred), it illustrates some of the concepts of generalized biodiversity indices introduced in previous chapters. At the same time it demonstrates the difficulties biodiversity conservation has to face. The results do not only depend on the biodiversity index used, but also on the phylogenetic tree or network the analysis is based on. Even in this small example, results differ between the species tree based on nuclear DNA, the species tree based on mitochondrial DNA and the network manually constructed. Thus, not only the biodiversity index to be used in a prioritization decision has to be chosen carefully, but also the phylogenetic tree or phylogenetic network the analysis is based on, in particular if the evolutionary history of a set of species is not fully understood.

8. Discussion

Due to limited financial means, biodiversity conservation has to prioritize the species to conserve. Existing approaches to prioritization are based on phylogenetic trees and use the *phylogenetic diversity* of subsets of taxa as a quantitative measure of biodiversity. *Phylogenetic diversity* serves as a basis for distinctiveness indices, in particular for the *Fair Proportion Index* and the *Shapley Value*, which rank species according to their contribution to overall biodiversity and thus, provide a simple prioritization criterion.

However, there are forms of non-treelike evolution, e.g. hybridization and horizontal gene transfer, which cannot be represented by phylogenetic trees. Thus, phylogenetic networks have come to the fore as a mathematical generalization of phylogenetic trees and are now an important concept in evolutionary biology, as they allow for the representation of reticulate (non-treelike) evolutionary events.

In this thesis, we provide a combination of both, i.e. we extend the concepts of *phylogenetic diversity* and biodiversity indices from phylogenetic trees to phylogenetic networks. For this purpose, we suggest a variety of approaches towards the use of *phylogenetic diversity*, the *Fair Proportion Index* and the *Shapley Value* in the context of hybridization networks.

In order to generalize the concept of *phylogenetic diversity* from trees to networks, we use three main principles: the calculation of spanning arborescences, the consideration of the (multi)set of phylogenetic trees displayed by a network and the construction of the *LSA tree* associated with a network. Similarly, we suggest to derive the *Fair Proportion Index* and the *Shapley Value* from the (multi)set of phylogenetic trees displayed by a network or the *LSA tree* associated with it. Additionally, we introduce a new index, the *Net Fair Proportion Index*, related to the *Fair Proportion Index* for phylogenetic trees, but defined on phylogenetic networks. Lastly, we suggest to derive the *Shapley Value* from any generalized definition of *phylogenetic diversity*.

All approaches have their specific advantages and drawbacks, in particular in ways of biological plausibility and computational feasibility. In theory, all approaches may face computational problems when applied to phylogenetic networks with a high number of reticulation nodes. In practice, this affects in particular the *embedded phylogenetic diversity* and any *embedded biodiversity index*, while the *LSA associated phylogenetic diversity* and *LSA associated diversity indices* seem to remain relatively unaffected by a high number of reticulation nodes. Under biological aspects, however, the *LSA associated phylogenetic diversity* and *LSA associated diversity indices* may be questionable, because the *LSA tree* associated with a network reduces the network to its most basic treelike content and thus, discards a lot of evolutionary information. The *embed-*

ded phylogenetic diversity, in particular $PD^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}$, and the *embedded biodiversity indices* $FP^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}$, $SV^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}$, $\widetilde{SV}^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}$ and $\widehat{SV}^{\varnothing_{hyb}}_{\mathsf{T}(\mathcal{N})}$, on the other hand, use most of the structure present in the network and seem biologically justified to use, in particular if we recall that evolution on the nucleotide level rather than the genome level is still treelike.

However, we strongly suggest to analyze the biological justification of all approaches in future research. In a second step, it might then be necessary to develop approximations for those definitions of generalized *phylogenetic diversity* and generalized biodiversity indices that seem particularly suitable for taxon prioritization in order to make them computationally feasible and applicable in practice.

In general, we suggest to use as much of the information provided by a phylogenetic network as possible when calculating the *phylogenetic diversity* of subsets of taxa or any biodiversity index. In particular, we suggest to incorporate hybridization probabilities, if they are given for a phylogenetic network. So far, our software tool `net_diversity.pl` only implements approaches independent of hybridization probabilities. Thus, the extension to approaches incorporating hybridization probabilities should be subject to further activities in this field.

Furthermore, we have solely focused on hybridization networks in this thesis and have not considered horizontal gene transfer networks or networks containing both hybridization and horizontal gene transfer events. Thus, the handling of horizontal gene transfer events should also be subject to further research. In principle, our approaches can be applied to horizontal gene transfer networks, but there might be a need for slight modifications. Recall that we have considered all reticulation edges, i.e. all edges directed into a reticulation node, to be unweighted (cf. Remarks on page 16). In case of horizontal gene transfer events, however, it might be necessary to distinguish between the actual reticulation edge, say e_{hgt}, directed into a reticulation node (i.e. the edge representing the horizontal gene transfer event) and the other edge, say e_t, directed into this node. Formally, the latter is also a reticulation edge, because it is directed into a reticulation node (cf. Definition 9), but we might want to treat it as a tree edge (cf. Figure 37). Subsequently the question arises, how to define the (multi)set of phylogenetic trees displayed by a horizontal gene transfer network. Following Definition 15, the (multi)set of phylogenetic trees of a network $\mathcal{N}$ can be obtained from $\mathcal{N}$ by all combinations of deleting one of the reticulation edges for each reticulation node and suppressing the resulting nodes of indegree 1 and outdegree 1. However, deleting the edge e_t (i.e. the edge that we might rather regard as a tree edge than as a reticulation edge) may result in trees, where the root has outdegree 1 (cf. Figure 37). This conflicts with our definition of a rooted phylogenetic X-tree (cf. Definition 2). Thus, either the definition of a rooted phylogenetic X-tree or the the definition of the (multi)set $\mathsf{T}(\mathcal{N})$

of trees displayed by a phylogenetic network might need to be modified when dealing with horizontal gene transfer events. Once this has been clarified, however, both the *embedded phylogenetic diversity* and the *embedded biodiversity indices* can be applied to horizontal gene transfer networks. All other approaches, e.g. the calculation of spanning arborescences or the construction of the *LSA tree* associated with a network should directly be applicable to horizontal gene transfer networks.

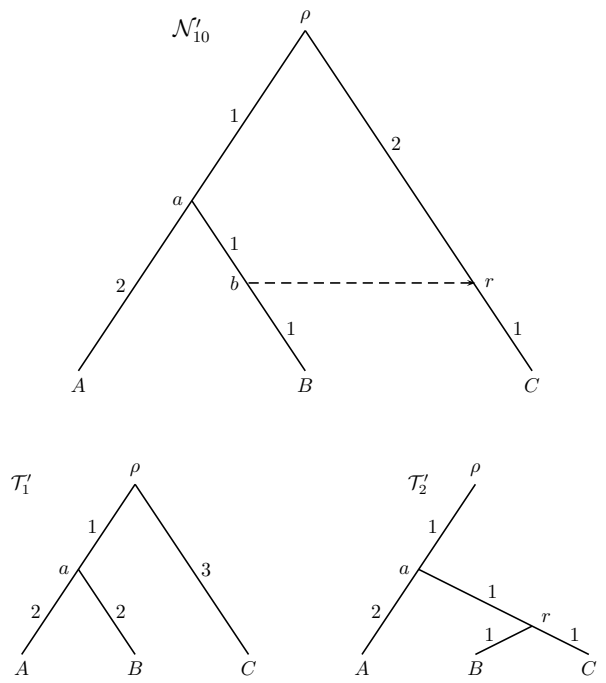


Fig. 37: Horizontal gene transfer network $\mathcal{N}'_{10}$ on $X = \{A, B, C\}$ and its embedded trees. Formally, the edges (b, r) and (ρ, r) are reticulation edges, because they are directed into the reticulation node r. However, we might want to treat the edge (ρ, r) as a tree edge. If we delete the edge (b, r) and suppress nodes of indegree 1 and outdegree 1, we retrieve the phylogenetic X-tree $\mathcal{T}'_1$. Deleting the edge (ρ, r) results in the tree $\mathcal{T}'_2$. Note that the root ρ has outdegree 1 in $\mathcal{T}'_2$ and thus, $\mathcal{T}'_2$ is not a rooted binary phylogenetic X-tree. Also note that the sum of branch lengths in $\mathcal{T}'_2$ is 6, while it is 8 in $\mathcal{N}'_{10}$ and $\mathcal{T}'_1$. It is therefore questionable whether $\mathcal{T}'_2$ should be regarded as displayed by $\mathcal{N}'_{10}$ or not.

In summary, our approaches provide an extension to existing prioritization tools in conservation biology and allow for the consideration of phylogenetic networks in prioritization decisions. This is of importance if the evolutionary history of a set of species is known to be non-treelike, and thus, cannot be represented by a phylogenetic tree. We remark, however, that further research concerning the biological plausibility of our approaches, their computational feasibility and the incorporation of horizontal gene transfer events is necessary before they can be put into practice.

References

[1] http://www.biologyreference.com/Ho-La/Hybridization-Plant.html. Online; accessed: 14-May-2016.

[2] http://search.cpan.org/~cjfields/BioPerl/Bio/PhyloNetwork.pm#eNewick_description. Online; accessed: 26-July-2016.

[3] https://wiki.rice.edu/confluence/download/attachments/5216841/RichNewick-2012-02-16.pdf?version=1&modificationDate=1330535426168&api=v2. Online; accessed: 25-September-2016.

[4] A. R. Amaral, J. A. Jackson, L. M. Möller, L. B. Beheregaray, and M. Manuela Coelho. Species tree of a recent radiation: The subfamily Delphininae (Cetacea, Mammalia). *Molecular Phylogenetics and Evolution*, 64(1): 243–253, Jul 2012. ISSN 1055-7903. doi: 10.1016/j.ympev.2012.04.004. URL http://dx.doi.org/10.1016/j.ympev.2012.04.004.

[5] A. R. Amaral, G. Lovewell, M. M. Coelho, G. Amato, and H. C. Rosenbaum. Hybrid Speciation in a Marine Mammal: The Clymene Dolphin (Stenella clymene). *PLoS ONE*, 9(1):e83645, Jan 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0083645. URL http://dx.doi.org/10.1371/journal.pone.0083645.

[6] G. Cardona, F. Rosselló, and G. Valiente. Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics*, 9(1):532, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-532. URL http://dx.doi.org/10.1186/1471-2105-9-532.

[7] G. Cardona, F. Rosselló, and G. Valiente. A perl package and an alignment tool for phylogenetic networks. *BMC Bioinformatics*, 9(1):175, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-175. URL http://dx.doi.org/10.1186/1471-2105-9-175.

[8] K. D. Cooper, T. J. Harvey, and K. Kennedy. A simple, fast dominance algorithm. URL https://www.cs.rice.edu/~keith/EMBED/dom.pdf.

[9] P. Cordue, S. Linz, and C. Semple. Phylogenetic Networks that Display a Tree Twice. *Bull Math Biol*, 76(10):2664–2679, Sep 2014. ISSN 1522-9602. doi: 10.1007/s11538-014-0032-x. URL http://dx.doi.org/10.1007/s11538-014-0032-x.

[10] D. Darriba, G. L. Taboada, R. Doallo, and D. Posada. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, 9(8):772–772, Jul 2012.

ISSN 1548-7105. doi: 10.1038/nmeth.2109. URL http://dx.doi.org/10.1038/nmeth.2109.

[11] D. P. Faith. Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61(1):1–10, 1992. ISSN 0006-3207. doi: 10.1016/0006-3207(92)91201-3. URL http://dx.doi.org/10.1016/0006-3207(92)91201-3.

[12] J. Felsenstein, J. Archie, W. Day, W. Maddison, C. Meacham, F. Rohlf, and D. Swofford. The newick tree format., 2000. URL http://evolution.genetics.washington.edu/phylip/newicktree.html.

[13] C. A. Floudas and P. M. Pardalos. *Encyclopedia of Optimization*. Springer, 2009. ISBN 0387747591.

[14] M. C. Fontaine, J. B. Pease, A. Steele, R. M. Waterhouse, D. E. Neafsey, I. V. Sharakhov, X. Jiang, A. B. Hall, F. Catteruccia, E. Kakani, and et al. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217):1258524–1258524, Nov 2014. ISSN 1095-9203. doi: 10.1126/science.1258524. URL http://dx.doi.org/10.1126/science.1258524.

[15] M. Fuchs and E. Y. Jin. Equality of Shapley value and fair proportion index in phylogenetic trees. *J Math Biol*, 71(5):1133–1147, Nov 2015. doi: 10.1007/s00285-014-0853-0. URL http://dx.doi.org/10.1007/s00285-014-0853-0.

[16] C.-J. Haake, A. Kashiwada, and F. E. Su. The Shapley value of phylogenetic trees. *J. Math. Biol.*, 56(4):479–497, Sep 2007. ISSN 1432-1416.

[17] M. Haase, M. Becker, and S. Zielske. Conflict of mitochondrial phylogeny and morphology-based classification in a pair of freshwater gastropods (caenogastropoda, truncatelloidea, tateidae) from New Caledonia. *ZooKeys*, 603:17–32, Jul 2016. ISSN 1313-2989. doi: 10.3897/zookeys.603.9144. URL http://dx.doi.org/10.3897/zookeys.603.9144.

[18] K. Hartmann. The equivalence of two phylogenetic biodiversity measures: the Shapley value and Fair Proportion index. *J Math Biol*, 67(5):1163–1170, Nov 2013. doi: 10.1007/s00285-012-0585-y. URL http://dx.doi.org/10.1007/s00285-012-0585-y.

[19] D. H. Huson and C. Scornavacca. Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. *Systematic Biology*, 61(6):1061–1067, Jul 2012. ISSN 1076-836X. doi: 10.1093/sysbio/sys062. URL http://dx.doi.org/10.1093/sysbio/sys062.

[20] D. H. Huson, R. Rupp, and C. Scornavacca. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press, New York, NY, USA, 2011. ISBN 0521755964, 9780521755962.

[21] F. K. Hwang, D. S. Richards, and P. Winter. *The Steiner Tree Problem (Annals of Discrete Mathematics)*. North-Holland, 1992. ISBN 044489098X.

[22] N. J. Isaac, S. T. Turvey, B. Collen, C. Waterman, and J. E. Baillie. Mammals on the EDGE: Conservation Priorities Based on Threat and Phylogeny. *PLoS ONE*, 2(3):e296, Mar 2007. ISSN 1932-6203. doi: 10.1371/journal.pone.0000296. URL `http://dx.doi.org/10.1371/journal.pone.0000296`.

[23] I. A. Kanj, L. Nakhleh, C. Than, and G. Xia. Seeing the trees and their branches in the network is hard. *Theoretical Computer Science*, 401(1-3):153–164, Jul 2008. ISSN 0304-3975. doi: 10.1016/j.tcs.2008.04.019. URL `http://dx.doi.org/10.1016/j.tcs.2008.04.019`.

[24] T. Lengauer and R. E. Tarjan. A fast algorithm for finding dominators in a flowgraph. *TOPLAS*, 1(1):121–141, Jan 1979. ISSN 0164-0925. doi: 10.1145/357062.357071. URL `http://dx.doi.org/10.1145/357062.357071`.

[25] S. Linz, K. S. John, and C. Semple. Counting Trees in a Phylogenetic Network is #P-Complete. *SIAM Journal on Computing*, 42(4):1768–1776, Jan 2013. ISSN 1095-7111. doi: 10.1137/12089394x. URL `http://dx.doi.org/10.1137/12089394X`.

[26] I. A. M. Marino, E. Riginella, M. Gristina, M. B. Rasotto, L. Zane, and C. Mazzoldi. Multiple paternity and hybridization in two smooth-hound sharks. *Scientific Reports*, 5:12919, Aug 2015. ISSN 2045-2322. doi: 10.1038/srep12919. URL `http://dx.doi.org/10.1038/srep12919`.

[27] B. Q. Minh, S. Klaere, and A. V. Haeseler. Phylogenetic Diversity on Split Networks. `http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=7A0B67BB82B5DF9B8B94963321A4F16B?doi=10.1.1.192.3523&rep=rep1&type=pdf`, 2007.

[28] J. A. T. Morgan, A. V. Harry, D. J. Welch, R. Street, J. White, P. T. Geraghty, W. G. Macbeth, A. Tobin, C. A. Simpfendorfer, and J. R. Ovenden. Detection of interspecies hybridisation in Chondrichthyes: hybrids and hybrid offspring between Australian (carcharhinus tilstoni) and common (c. limbatus) blacktip shark found in an Australian fishery. *Conservation Genetics*, 13(2):455–463, Dec 2011. ISSN

1572-9737. doi: 10.1007/s10592-011-0298-6. URL http://dx.doi.org/10.1007/s10592-011-0298-6.

[29] F. Ronquist, M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Hohna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology*, 61(3):539–542, Feb 2012. ISSN 1076-836X. doi: 10.1093/sysbio/sys029. URL http://dx.doi.org/10.1093/sysbio/sys029.

[30] J. E. Stajich. The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research*, 12(10):1611–1618, Oct 2002. ISSN 1088-9051. doi: 10.1101/gr.361602.

[31] B. Sziklai, T. Fleiner, and T. Solymosi. On the Core of Directed Acyclic Graph Games. IEHAS Discussion Papers 1418, Institute of Economics, Centre for Economic and Regional Studies, Hungarian Academy of Sciences, 2014. URL http://EconPapers.repec.org/RePEc:has:discpr:1418.

[32] C. Than, D. Ruths, and L. Nakhleh. Phylonet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, 9 (1):322, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-322. URL http://dx.doi.org/10.1186/1471-2105-9-322.

[33] L. Volkmann, I. Martyn, V. Moulton, A. Spillner, and A. O. Mooers. Prioritizing Populations for Conservation Using Phylogenetic Networks. *PLoS ONE*, 9(2): e88945, Feb 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0088945. URL http://dx.doi.org/10.1371/journal.pone.0088945.

[34] R. A. Vos, J. Caravas, K. Hartmann, M. A. Jensen, and C. Miller. BIO::Phylo-phyloinformatic analysis using perl. *BMC bioinformatics*, 12(1):63, 2011.

[35] D. Wen, Y. Yu, M. W. Hahn, and L. Nakhleh. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Mol Ecol*, 25(11):2361–2372, Mar 2016. ISSN 0962-1083. doi: 10.1111/mec.13544. URL http://dx.doi.org/10.1111/mec.13544.

[36] K. Wicke. The Shapley Value and the Fair Proportion Index as measures of Biodiversity – Analysis, Comparison and Computation. Bachelor thesis, University of Greifswald, Germany, Oct. 2014.

[37] K. Wicke and M. Fischer. Comparing the rankings obtained from two biodiversity indices: the Fair Proportion Index and the Shapley Value. July 2015. URL https://arxiv.org/abs/1507.08620.

[38] S. Zielske and M. Haase. Molecular phylogeny and a modified approach of character-based barcoding refining the taxonomy of New Caledonian freshwater gastropods (caenogastropoda, truncatelloidea, tateidae). *Molecular Phylogenetics and Evolution*, 89:171–181, Aug 2015. ISSN 1055-7903. doi: 10.1016/j.ympev.2015.04.020. URL `http://dx.doi.org/10.1016/j.ympev.2015.04.020`.

A. Table of contents of the CD

- `Delphinidae`
  - `1_Phylogenetic_Analysis`
    - `cytochrome_b.nex` ..... alignment for the cytochrome *b* gene
    - `dataset_Amaral2012.nex` ..... dolphin data set [4]
    - `mrBayes_cytochrome_b.tre` ..... Bayesian phylogenetic species tree for the cytochrome *b* gene
    - `nuclear_species_tree_Amaral2014.tre` ..... nuclear phylogenetic species tree presented in [5]
  - `2_Dendroscope_Network_Inference`
    - `dendroscope_hybridization_networks.tre` .. hybridization networks inferred with Dendroscope [19] in the extended Newick format
    - `dendroscope_hybridization_networks.pdf` ..... images of the hybridization networks
    - `nuclear_and_mitochondrial_trees.tre` .. nuclear and mitochondrial species tree in the Newick tree format (input for Dendroscope [19])
  - `3_Hybridization_Network_Manually`
    - `net_embedded_trees.tre` ..... phylogenetic trees displayed by the dolphin hybridization network in the extended Newick format
    - `net_lsa_tree.tre` ..... *LSA tree* associated with the dolphin hybridization network in the extended Newick format
    - `net_manually.tre` ..... dolphin hybridization network in the extended Newick format
    - `Shapley_Results.txt` ..... *original Shapley Values* for the dolphin hybridization network
- `PDF`
  - `MasterThesis_KristinaWicke.pdf` ..... electronic copy of this thesis
- `Program`
  - `net_diversity.pl` ..... Perl script introduced in this thesis
  - `network_N2.txt` ..... example: network $\mathcal{N}_2$ in the extended Newick format
  - `network_N8.txt` ..... example: network $\mathcal{N}_8$ in the extended Newick format

- `README.txt` ..... README file for `net_diversity.pl`

- `RandomNetworks`
  - `10Taxa`
    - `Net_10Taxa_0Hybrids.txt` ..... random phylogenetic network with 10 leaves and 0 reticulation nodes in the extended Newick format
    - `Net_10Taxa_5Hybrids.txt` ..... random phylogenetic network with 10 leaves and 5 reticulation nodes in the extended Newick format
    - `Net_10Taxa_9Hybrids.txt` ..... random phylogenetic network with 10 leaves and 9 reticulation nodes in the extended Newick format
  - `20Taxa`
    - `Net_20Taxa_0Hybrids.txt` ..... random phylogenetic network with 20 leaves and 0 reticulation nodes in the extended Newick format
    - `Net_20Taxa_5Hybrids.txt` ..... random phylogenetic network with 20 leaves and 5 reticulation nodes in the extended Newick format
    - `Net_20Taxa_10Hybrids.txt` .... random phylogenetic network with 20 leaves and 10 reticulation nodes in the extended Newick format
    - `Net_20Taxa_15Hybrids.txt` .... random phylogenetic network with 20 leaves and 15 reticulation nodes in the extended Newick format
  - `50Taxa`
    - `Net_50Taxa_49Hybrids.txt` .... random phylogenetic network with 50 leaves and 49 reticulation nodes in the extended Newick format
  - `75Taxa`
    - `Net_75Taxa_74Hybrids.txt` .... random phylogenetic network with 75 leaves and 74 reticulation nodes in the extended Newick format
  - `100Taxa`
    - `Net_100Taxa_50Hybrids.txt` . random phylogenetic network with 100 leaves and 50 reticulation nodes in the extended Newick format
  - `network_generator.pl` . Perl script used to generate random phylogenetic networks
  - `tree_generator.pl` ..... Perl script used to generate random phylogenetic trees

Declaration of Authorship

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references.
This paper was not previously presented to another examination board and has not been published.

Greifswald, 8th November 2016