Distribution and asymptotic behavior of the phylogenetic transfer distance

Miraine Dávila Felipe[1], Jean-Baka Domelevo Entfellner[2], Frédéric Lemoine[1,3], Jakub Truszkowski[4], and Olivier Gascuel [*1,4]

[1] *Unité Bioinformatique Evolutive / C3BI USR 3756, Institut Pasteur & CNRS, Paris, France*

[2] *Biosciences eastern and central Africa (BecA-ILRI Hub), International Livestock Research Institute, PO Box 30709, Nairobi 00100, Kenya*

[3] *Hub Bioinformatique et Biostatistique, C3BI USR 3756, Institut Pasteur & CNRS, Paris, France*

[4] *Méthodes et Algorithmes pour la Bioinformatique, IBC - LIRMM UMR 5506, Université de Montpellier & CNRS, Montpellier, France*

July 13, 2018

Abstract

The *transfer distance* (TD) was introduced in the classification framework and studied in the context of phylogenetic tree matching. Recently, Lemoine et al. (2018) showed that TD can be a powerful tool to assess the branch support of phylogenies with large data sets, thus providing a relevant alternative to Felsenstein's bootstrap. This distance allows a *reference branch* β in a reference tree $\mathcal{T}$ to be compared to a branch b from another tree T, both on the same set of n taxa. The TD between these branches is the number of taxa that must be transferred from one side of b to the other in order to obtain β. By taking the minimum TD from β to all branches in T we define the *transfer index*, denoted by $\phi(\beta, T)$, measuring the degree of agreement of β with T. Let us consider a reference branch β having p tips on its light side and define the *transfer support* (TS) as $1 - \phi(\beta, T)/(p-1)$. The aim of this article is to provide evidence that $p-1$ is a meaningful normalization constant in the definition of TS, and measure the statistical significance of TS, assuming that β is compared to a tree T drawn according to a null model. We obtain several results that shed light on these questions in a number of settings. In particular, we study the asymptotic behavior of TS when n tends to ∞, and fully characterize the distribution of ϕ when T is a caterpillar tree.

*Corresponding author: olivier.gascuel@pasteur.fr

1 Introduction

The *transfer distance* or *R-distance* was introduced in the classification framework by Day [5] and Régnier [22], as a measure of (dis)similarity between partitions of a set. It is defined as the minimum number of elements that need to be removed from their original class or transferred from one class to another, in order to transform one partition into the other. This distance possesses some desirable properties, for example its low computational cost in comparison with other metrics, as established by Day [5]. Charon, Denœud, Guénoche, and Hudry [4] studied other characteristics of this distance such as the maximum transfer distance that can be obtained when comparing two partitions with a fixed, but possibly different, number of classes. As highlighted by Denœud [7], it proves challenging to study the theoretical properties of the transfer distance, so the author proposed an experimental analysis of the transfer distance using simulations, to discuss its interpretation and approximate its distribution and mean when considering pairs of random partitions of the same set.

The interest in using the transfer distance to compare phylogenetic trees started with a seminal paper by Day in 1985 [6]. In the field of computational biology, problems involving tree comparison have remained a major challenge for many years. A common concern with most of these problems is to define a suitable metric on trees. The transfer distance is a measure to compare bipartitions, and a phylogenetic tree is unambiguously defined by the set of bipartitions induced by its branches. Then, a logical question to ask is whether we can define a metric on trees based on this transfer distance on bipartitions. There is not a unique way of defining such a metric and several authors have worked on the subject. For instance, in [6], the author proposes several algorithms and methods to solve related tree problems, in particular the construction of the consensus of a set of trees. As discussed in [6], this task requires the optimization of a consensus index, which can be defined using the transfer distance or other metrics, such as the well-known Robinson-Foulds (RF) metric [21]. The latter is probably the most widely used distance between trees and is defined as the number of bipartitions belonging to one tree but not to the other. However, the RF metric is known to have several drawbacks, including its lack of robustness, since it is highly sensitive to certain small tree changes, as pointed out by Lin, Rajan, and Moret [19] and Bogdanowicz and Giaro [3].

In another study, Boc, Philippe, and Makarenkov [2] proposed to optimize a tree comparison index that can also be based on metrics, including RF and the transfer distance. Having set a goal to detect accurately horizontal gene transfer events, the authors showed that the version of their algorithm relying on the transfer distance provides the best results when searching for an optimal scenario of Subtree Pruning and Regrafting (SPR) moves needed to transform a gene tree into a species tree. However, the

transfer-based dissimilarity defined in [2] is not a metric, since it violates the triangle inequality in some cases [2, p.197, Prop. 1]. Lin, Rajan, and Moret [19] addressed this problem by proposing a different distance measure based on minimum-cost matching between the two sets of splits induced by both trees. For that metric, also relying on the transfer distance, the triangle inequality holds. Additionally, the computational and statistical properties of this new distance are studied in [19], where the authors provide a low-polynomial time algorithm, establishing its robustness through statistical testing and demonstrating its usefulness in clustering trees.

Recently, we proposed a new bootstrap method for large phylogenetic trees, relying on branch comparisons based on the transfer distance [18]. The aim of that study was to use a more fine-grained measure for the presence of a branch in a tree, rather than the binary values used in Felsenstein's classical bootstrap technique [11]. We compared a *reference branch* β in a *reference tree* $\mathcal{T}$ to another tree T, typically a bootstrap tree, by taking the minimum of the transfer distance from β to any branch b in T, which we called the *transfer index* and denoted by $\phi(\beta, T)$. Next, we averaged the values of ϕ over a set of bootstrap trees, obtaining, after appropriate normalization, the so-called *transfer bootstrap expectation* (TBE). We explored the behavior of TBE as a measure of support for the branches of a phylogenetic tree, compared to that of Felsenstein's support (FS). In a number of experiments using both real and simulated data, we found that TBE outperformed FS. This was particularly noticeable for deep branches and large values of n, where FS often failed to detect the phylogenetic signal in the trees. In view of those results, TBE shows promise as a useful tool in phylogenetic analysis. In [18], we studied and discussed several of its mathematical properties, but there is still need for further work so that the transfer index and TBE are fully understood. The main motivation for the present work is therefore to study the properties of the transfer index and support, assess their statistical significance, and give analytic expressions for TBE when the reference branch is compared to a tree T drawn randomly according to some null model.

To be more specific about the results obtained here, fix $n \geq 4$ and let us consider phylogenetic trees on a set X of n taxa. To distinguish the two sides of the reference bipartition β, we say that its *light side* contains $p \geq 2$ taxa while its *heavy side* has $n - p \geq p$ taxa. The TBE we proposed in [18] is actually the average, over all the bootstrap trees, of the *transfer support* function (TS), which is defined as $1 - \frac{\phi(\beta,T)}{p-1}$. It is not hard to see that $\phi(\beta, T) \leq p - 1$ and thus the TS function takes values on the interval $[0, 1]$. Notice that we deliberately exclude here cases with $n = 2, 3$, and $p = 1$, which result in trivial bipartitions.

First, let us reproduce the results obtained from computer simulations in [18], where we looked at four different models for the tree T. More precisely, we considered (1) two simple models for the topology of phylogenetic trees:

caterpillar trees and perfectly balanced binary trees; and (2) two classic null models for speciation: the Proportional to Distinguishable Arrangements (PDA) and Yule-Harding models (see Section 2 for further details). For each of these models, we performed simulations for different values of n (128, 256, 512, and 1,024) and all the possible values of p for each n (i.e. $2 \leq p \leq \lfloor n/2 \rfloor$). Then, for each value of (n, p) and for each model, we simulated a set of reference bipartitions and a set of trees to be compared to these reference bipartitions. The results are displayed in Fig. 1, where we plotted the TS values thus obtained against the different values of p. This figure shows striking evidence that, on average, TS stays close to 0 with random trees, so the transfer index stays close to its upper bound $p - 1$ for all p. Additionally, we observe that the maximum value of the TS attained over all possible values of p seems to be obtained at $p = \lfloor n/2 \rfloor$ and to decrease when n increases. Moreover, this asymptotic behavior seems to be independent of the topology/model considered for the tree T. These initial observations motivate the questions we address here.

The intuition provided by computer simulations led us to obtain several theoretical results that justify what we observed in Fig. 1 and the simple normalization by $p - 1$ used in [18]. We also formulate a set of conjectures that are discussed in Section 6. Our first result consists in the characterization of the asymptotic behavior of the transfer index $\phi(\beta, T)$ for a random bipartition β of the set X and any tree T. More precisely, we prove that the transfer index converges in probability to $p - 1$ when p is fixed (but see below) and n tends to ∞. The proof relies on the comparison of the transfer index with the parsimony score of a binary character and the use of a result from Steel and Penny [25]. We then use concentration inequalities to characterize the asymptotic behavior of the transfer index when p grows, depending on n (e.g. when $p \to \lfloor n/2 \rfloor$). Lastly, when T is a caterpillar tree, we fully characterize the probability distribution of the transfer index based on a one-to-one correspondence between these trees and North-East (NE) lattice paths, a common technique for counting combinatorial objects [20]. All of these results show that $p - 1$ is the appropriate normalization constant for the TS proposed in [18] and that this support has the expected behavior, which is that, in the absence of phylogenetic signal in the tree T regarding the reference branch β, the TS is close to 0, especially for large trees.

The paper is organized as follows. In Section 2, we give the main definitions and properties of the concepts described earlier. Section 3 is devoted to the results concerning the parsimony score, and Section 4 presents the asymptotic results using concentration inequalities. Details on the specific case of the caterpillar tree are given in Section 5.

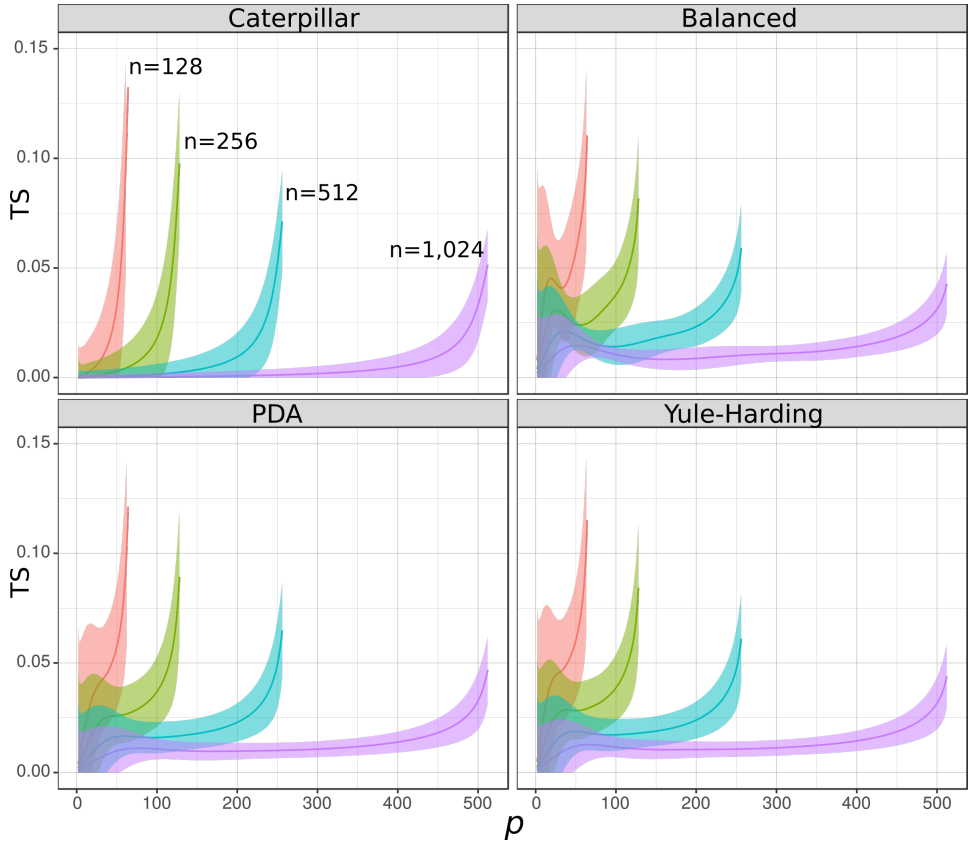


Figure 1: Simulations of TS for random trees. The four panels correspond to the four different models considered: caterpillar, totally balanced, PDA, and Yule-Harding. For each of these models, we considered four different values of n and used the following color codes to distinguish them: 128 (red), 256 (green), 512 (blue), and 1,024 (purple). For each model and for each value of n and p (from 2 to $\lfloor n/2 \rfloor$), we simulated 100 reference bipartitions and 1,000 trees to be compared to the reference bipartition. We plotted the mean of the $100 \times 1,000$ TS values obtained for each n and for each model (bold lines) and the standard deviation (shaded areas), against p.

2 Preliminaries

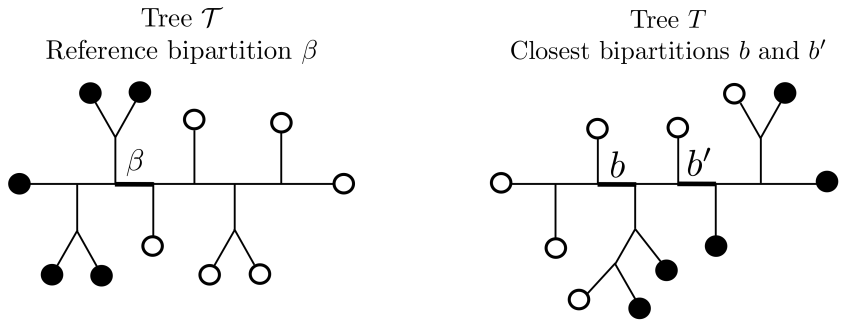
In this section, we give the main definitions and general properties on phylogenetic trees that are needed for the rest of the paper. We refer to [23] for an extensive mathematical treatment of this subject.

Let us fix $n \geq 4$ and X, a set of n taxa. We consider phylogenetic trees on X, that is trees whose leaves are mapped one-to-one to X. These trees are called *phylogenetic X-trees* or simply phylogenies. For simplicity of notation, we shall always take $X = \{1, 2, \ldots, n\}$. Denote by $\mathrm{UB}(n)$ the set

of all unrooted *binary* phylogenetic trees (every interior vertex has degree 3) on n leaves. For a phylogenetic tree T, we use $\mathcal{E}(T)$, $\mathcal{V}(T)$ to denote respectively the set of edges and the set of vertices of the tree.

For any X-tree T, a branch $b \in \mathcal{E}(T)$ can be encoded in several equivalent ways, that we will use indistinctly depending on the context. First, any branch b defines a bipartition (or split), and we can associate b to a vector $v(b)$ in $\{0,1\}^n$ by assigning the same number (e.g. 0) to all the elements on the same side of the split induced by this branch. Notice that b is also encoded by $\overline{v}(b)$, the negation of v (i.e. the 0 values are turned into 1 and vice versa). Likewise, we can identify a bipartition, with a *bicoloration* of the leaves, that is a function that assigns one of two colors (black = **B** or white = **W**) to each leaf label. Notice however, that we can consider a bipartition or a bicoloration on a tree that does not correspond to any branch in this tree. To make the distinction, we say $b \in \mathcal{E}(T)$ for the bipartitions induced by branches on the tree T, and we use $\mathcal{X} \coloneqq \{f : X \to \{\mathbf{B}, \mathbf{W}\}\}$ to denote the set of all possible bicolorations of the tips of T. Then, $b \in \mathcal{X}$ does not necessarily correspond to a branch in T, but to a bicoloration of its tips.

We will leverage the visual aspect of the two-color representation and, throughout the rest of the article, we associate the p taxa of the light side in the reference bipartition with the black color **B** and the $n-p$ ($\geq p$) taxa of the heavy side with the white color **W** (Fig. 2, left). The set of bicolorations satisfying that $|\{i \in X : f(i) = \mathbf{B}\}| = p$ will be denoted by $\mathcal{X}_p$. A tree T endowed with a bicoloration is called a bicolored tree.



(a) (b)

Figure 2: (a) An example of reference branch β dividing X in $p = 5$ black tips and $n - p = 6$ white tips, respectively. (b) Another X-tree T in which the closest bipartitions to β are b and b', both giving a transfer distance $\delta(\beta, b) = \delta(\beta, b') = 3$, and thus $\phi(\beta, T) = 3$ as no other branch in T is closer to β.

As described in the Introduction, the transfer distance is used to compare a branch β in the reference phylogenetic tree topology $\mathcal{T}$, to a second branch

b in another tree topology T, both on the same taxa set X. This distance can easily be defined using the Hamming distance $H(\cdot,\cdot)$ between two vectors of equal size.

Definition 1 (Transfer distance)**.**

$$\delta(\beta, b) \coloneqq \min\{H\left(v(\beta), v(b)\right), H\left(\overline{v}(\beta), v(b)\right)\}.$$

Based on this definition, notice that $\delta(\beta, b) = 0$ if and only if $v(\beta)$ and $v(b)$ define the same bipartition. To measure the degree of presence of β in T, we define the *transfer index*, denoted by $\phi(\beta, T)$, which is the minimum of the transfer distance over all branches in T [18].

Definition 2 (Transfer index)**.**

$$\phi\left(\beta, T\right) \coloneqq \min_{b \in \mathcal{E}(T)} \delta(\beta, b).$$

As mentioned before, we are interested in the case where the reference tree $\mathcal{T}$ and a branch β on this tree are fixed. The core idea in [18] is to measure the presence of this reference branch in a set of bootstrap trees by using the following *transfer support* function.

Definition 3 (Transfer support)**.**

$$\mathrm{TS}\left(\beta, T\right) \coloneqq 1 - \frac{\phi(\beta, T)}{p-1}.$$

The transfer index and support functions satisfy simple properties that we can find in [18], and that are included here for completeness:

(i) $\phi\left(\beta, T\right) = 0 \iff \beta \in T$,

(ii) $\phi\left(\beta, T\right) \in [0, p-1]$ or equivalently $\mathrm{TS}\left(\beta, T\right) \in [0, 1]$.

The first statement can be deduced directly from the definition of $\delta(\beta, b)$, which is 0 if and only if $v(\beta) = v(b)$ or $\overline{v}(\beta) = v(b)$. Thus, $\phi\left(\beta, T\right) = 0$ if and only if we can find the bipartition induced by β in T. Moreover, we say that a bipartition is trivial when it has a single leaf on one side, and the remaining $n-1$ leaves on the other. For a trivial bipartition b defined by a taxon that belongs to the light side in β, we obtain that $\delta(\beta, b) = p-1$, and since ϕ is defined as the minimum taken over all possible branches in T, we obtain the statement (ii).

Null models

The aim of this study is to characterize the distribution and the asymptotic behavior of the transfer index and transfer support when the reference bipartition β is compared to a binary phylogenetic X-tree T that follows a

certain null model. We are interested mainly in unrooted trees, but it should be noted that the existence of a root has no influence on transfer distance values: both branches adjacent to the root define the same bipartition.

There are two ways to define the probabilistic models we are considering. First, we can suppose that we have a fixed bicoloration $\chi_p \in \mathcal{X}_p$ and that we draw a tree T randomly from $\mathrm{UB}(n)$, following some specific (probabilistic) model. Another way is to consider that the tree is fixed and a bicoloration of its tips is uniformly chosen from $\mathcal{X}_p$. In the first case, an interesting question is to consider the probabilistic models that are most commonly used in the field of phylogenetics [14], such as the Yule-Harding or PDA models. On the other hand, for a fixed tree, a natural question is to look at the two extreme cases for the topology regarding balance. The most imbalanced tree is called the caterpillar tree and can be defined as a binary phylogenetic tree for which the induced subtree on the interior vertices forms a path graph (if the tree is rooted, then the root is at one end of the path). On the other side, we have perfectly balanced trees, that is rooted binary phylogenetic trees with $n = 2^h$ leaves (for some $h \in \mathbb{N}$), each of which is at a distance of exactly h edges from the root. We refer to [23, 24] for further details on these tree models.

As explained in the Introduction, we performed computer simulations for these four models to exhibit their asymptotic properties (Fig. 1). We observe that the asymptotic behavior of the TS seems to be independent of the model considered, which we explain in the following sections. Then, a full theoretical treatment is carried out for the caterpillar model.

3 Comparing the transfer index to the parsimony score

We are now interested in comparing the transfer index to the widely used parsimony score introduced by Farris [9], Fitch [12], and Hartigan [15]. We then use the result to obtain a first characterization of the asymptotic behavior of the transfer index.

Definition 4 (Parsimony score)**.** Consider a phylogenetic tree T and a bicoloration of its tips $\chi \in \mathcal{X}$. Consider an extension of this bicoloration to all the nodes in T and denote it by $\overline{\chi}$ (each internal node is also assigned one of the two colors). Define

$$\mathrm{ps}\,(\chi, T) := \min_{\overline{\chi}} \sum_{b \in \mathcal{E}(T)} \mathrm{diff}(b, \overline{\chi}),$$

where $\mathrm{diff}(b, \overline{\chi})$ is 0 if both nodes connected by b have the same color in the extension $\overline{\chi}$, and 1 otherwise.

By using a simple argument, one can prove the following result from [18], given here for the sake of completeness.

Lemma 5. *For any given X-tree T and any bicoloration $\chi \in \mathcal{X}$, we have that*

$$\mathrm{ps}\,(\chi, T) \leq \phi\,(\chi, T) + 1. \tag{1}$$

Proof. Consider a branch b in T, and suppose it defines a bipartition having respectively $B_l(b), W_l(b), B_h(b), W_h(b)$ black and white tips on its light and heavy sides. We know from the definition that

$$\delta\,(\chi, b) = \min\{W_l + B_h, W_h + B_l\}. \tag{2}$$

Suppose, without loss of generality, that the minimum is $W_l + B_h$. We will now look at the parsimony score for this bicoloration. We put a character change at each of the W_l white leaves on the light side of b and at each of the B_h black leaves on its heavy side (*parents* take the opposite color). Then the internal nodes on the light side are colored in black and those on the heavy side in white. The number of color changes of this extension is $W_l + B_h + 1$ because we have to add an extra change at branch b. Since the parsimony score of the bicoloration is the minimum taken over all the possible extensions, we have that

$$\mathrm{ps}\,(\chi, T) \leq W_l + B_h + 1 = \delta\,(\chi, b) + 1.$$

Since this is true for any branch $b \in T$, we obtain the announced result. □

3.1 Asymptotic results for fixed p

In this subsection, we use inequality (1) between the parsimony score and the transfer index to establish that the transfer index converges to $p-1$ when p is fixed and n grows to infinity. Let us consider a random bicoloration χ_p from $\mathcal{X}_p$. Let T be any binary tree topology with n tips colored by χ_p. The larger n, the more dispersed the black tips in T, and the higher the probability that the parsimony score is equal to p and the transfer index to $p-1$. This is formalized as follows.

Proposition 6. *Let T be any binary tree topology with n tips, and χ_p be a random bicoloration of the tips of T, uniformly chosen from $\mathcal{X}_p$. We have*

$$\mathbb{P}\,(\phi\,(\chi_p, T) = p-1) \geq \mathbb{P}\,(\mathrm{ps}\,(\chi_p, T) = p) \geq 1 - 4n \times \frac{p(p-1)}{n(n-1)}.$$

Proof. The first inequality is an obvious consequence of inequality (1) and upper bound $p-1$ on the transfer index. To demonstrate the second inequality, we use a result from Steel and Penny [25, Prop. 9.4.1], stating that

if the color (state) changes in a tree are rare enough that any two edges with changes are separated by at least three edges with no changes, then the parsimony score is guaranteed to coincide exactly with the number of changes within the tree. Since χ_p contains p black tips, if we color all internal nodes white, we will have p changes on the external edges leading to the black tips. If the number of edges separating any pair of two black tips is at least 5, then the parsimony score $\mathrm{ps}\,(\chi_p, T)$ is equal to p, as all changes are separated by at least 3 internal edges with white vertices at both ends. Thus, we have:

$$\begin{aligned}\mathbb{P}\,(\mathrm{ps}\,(\chi_p, T) = p) &\geq \mathbb{P}\,(\text{any pair of black tips is at distance} \geq 5)\\ &\geq 1 - \mathbb{P}\,(\text{at least one pair of black tips is at distance} \leq 4)\,.\end{aligned}$$

For any tip in the (binary) tree T, the number of tips that are at most 4-edge distant is at most 8, and thus the number of pairs of tips at a distance of 4 edges or less is at most $4n$. The probability of drawing such a pair of tips (among $n(n-1)/2$) is

$$\mathbb{P}\,(\text{a pair of tips is at distance} \leq 4) \leq \frac{4n}{n(n-1)/2}.$$

As we have $p(p-1)/2$ pairs of black tips, we obtain the desired inequality by the union bound. □

This result is valid for any topology T as long as the bicoloration χ_p is uniformly distributed in the set $\mathcal{X}_p$. It has the following immediate consequences.

Corollary 7. *For any tree T, and any χ_p uniformly distributed in the set $\mathcal{X}_p$, we have that,*

- *when p is fixed, the transfer index $\phi(\chi_p, T)$ converges in probability to $p-1$ when $n \to \infty$;*
- *when $p = o(\sqrt{n})$, we have that $\phi(\chi_p, T) - (p-1)$ converges in probability to 0 when $n \to \infty$.*

4 Behavior of the transfer distance when p grows with n

In the previous section, we showed that, when n tends to infinity, the transfer index converges in probability to $p-1$ for fixed p, and TS converges to 0 when p grows slowly as $o(\sqrt{n})$. However, simulations in Fig. 1 suggest that the transfer index also behaves in a similar manner for larger values of p relative to n. For example, for all null models when $p = \lfloor n/2 \rfloor$, the expected value

of TS is larger than 0.1 with $n = 128$, but lower than 0.05 with $n = 1,024$. In this section, we will show that for "all values of p", the distribution of the transfer index is *concentrated* around $p-1$, meaning that the probability of the transfer index being "far away" from $p-1$ vanishes as n grows. This explains what we observe in our simulations and motivates the use of $p-1$ as the normalization term in the definition of TS.

The results we obtain in this section are based on concentration inequalities. More precisely, we make use of the well-known Chernoff-Hoeffding bounds for sums of independent random variables, as stated by Dubhashi and Panconesi [8]. In his original paper, Hoeffding [17] proved that these inequalities also hold for sums of variables obtained by sampling without replacement, which is the case of interest here. The following lemma is a direct consequence of the results in [17] and [8].

Lemma (Chernoff-Hoeffding bound)**.** *Let $X = \sum_{i=1}^{m} X_i$ where X_i are drawn without replacement from a multiset containing elements between 0 and 1. Then, for any $r > 0$*

$$\mathbb{P}(X \geq (1+r)\mathbf{E}[X]) \leq \left(\frac{e^r}{(1+r)^{(1+r)}} \right)^{\mathbf{E}[X]} \tag{3}$$

and for any $t > 0$, we have

$$\mathbb{P}(X \geq \mathbf{E}[X] + t) \leq e^{\frac{-2t^2}{m}}. \tag{4}$$

We can now state the main theorem of this section.

Theorem 8. *Let T be any binary tree topology with n tips, and let χ_p be a bicoloration of the tips in T chosen uniformly at random from $\mathcal{X}_p$. Then, there exists $N \in \mathbb{N}$, s.t. for all $n \geq N$, with probability at least $1 - O(\frac{1}{n})$,*

1. *if $p = O(n^\alpha)$ for some $0 < \alpha < 1$, then $\phi(\chi_p, T) \geq p - C$ for some constant C;*
2. *if $p = cn + o(n)$ for some $0 < c < 1/2$, then $\phi(\chi_p, T) \geq p - C \log n$ for some constant C;*
3. *if $p = \frac{1}{2}n - o(n)$, then $\phi(\chi_p, T) \geq p - C\sqrt{n \log n}$ for some constant C.*

These three cases correspond to different growth rates of p, from the slowest (1) to the fastest (3). Case 1 is already partly covered for $0 < \alpha < 1/2$ by Proposition 6 and Remark 7, which imply that for these values of p,

$$\mathbb{P}\left(\phi(\chi_p, T) = p - 1\right) \geq 1 - O\left(\frac{1}{n^{1-2\alpha}}\right).$$

Case 2 corresponds to what we observe in Fig. 1 (see also Extended Data Fig. 1 in [18]), where for any given ratio p/n (e.g. $p/n = 1/4$), the expected

value of the TS decreases when n increases. In Case 3, p is as large as possible, and the difference between ϕ and $p-1$ is the largest among all three cases, as expected. Note that the bound in Case 3 also holds when $p = n/2$ when n is even and $p = n/2 - 1$ when n is odd.

Corollary 9. *For any tree T, any χ_p uniformly distributed in the set $\mathcal{X}_p$, and any p that grows with n as in cases 1,2, and 3, the transfer support $TS(\chi_p, T)$ converges in probability to 0 when $n \to \infty$.*

Proof of Theorem 8. Consider a bipartition $b \in T$ and let $s \in \{l, h\}$ denote the light/heavy side of b. Let $q_s(b)$ be the number of taxa on side s of b and $B_s(b)$ be the random variable corresponding to the number of black taxa on side s of b. The transfer distance between b and the bicoloration χ_p can be written as

$$\delta(\chi_p, b) = \min\left\{p + q_l(b) - 2B_l(b), p + q_h(b) - 2B_h(b)\right\}.$$

Consequently, we can write the transfer index as

$$\phi(\chi_p, T) = \min_{b \in T, s \in \{l,h\}} p + q_s(b) - 2B_s(b).$$

For any $1 < u < p$, define $\mathcal{B}_u = \{b \in T : q_l(b) \geq u\}$, the set of bipartitions in T with at least u tips on both sides. We are interested in the set $\mathcal{B}_u$ since only the bipartitions in this set can give $\delta(\chi_p, b) \leq p - u$. This statement derives from some simple arguments based on the definition of the transfer distance. Consider any $b' \notin \mathcal{B}_u$, we necessarily have $q_l(b') < u < p < q_{b'h}$, $0 \leq B_l(b') \leq q_l(b')$, and $0 \leq B_h(b') \leq p$, which implies that

$$\begin{aligned} p + q_l(b') - 2B_l(b') &\geq p - q_l(b') > p - u, \\ p + q_h(b') - 2B_h(b') &\geq q_h(b') - p > p - q_l(b') > p - u. \end{aligned}$$

As a consequence, we can bound the tail probability of the transfer index using the union bound over all bipartitions in $\mathcal{B}_u$, that is

$$\begin{aligned} \mathbb{P}\left(\phi(\chi_p, T) \leq p - u\right) &\leq \sum_{b \in \mathcal{B}_u} \mathbb{P}\left(\delta(\chi_p, b) \leq p - u\right) \\ &\leq \sum_{b \in \mathcal{B}_u, s \in \{l,h\}} \mathbb{P}\left(B_s(b) \geq (q_s(b) + u)/2\right). \end{aligned} \tag{5}$$

We will now derive a bound on each of the elements of the above sum. First, notice that every $B_s(b)$ can be written as

$$B_s(b) = \sum_{i=1}^{q_s(b)} X_{bsi},$$

where $X_{bsi} = 1$ if the i-th leaf on side s of b is colored black and 0 otherwise. Thus, $B_s(b)$ follows a hypergeometric distribution and we have

$$\mathbb{E}\left[B_s(b)\right] = \frac{pq_s(b)}{n}.$$

Moreover, Chernoff-Hoeffding inequalities (3) and (4) apply to these hypergeometric variables and enable us to derive the appropriate bounds.

We now must consider three cases depending on the growth rate of p with respect to n.

Case 1. $p = O(n^\alpha)$ for some $\alpha < 1$.

Applying the bound in (3) with $r = n\left(q_s(b) + u\right) / \left(2pq_s(b)\right) - 1$, we get

$$\begin{aligned}
\mathbb{P}\left(B_s(b) \geq \frac{q_s(b)+u}{2}\right) &\leq \left(\frac{e^{\frac{n(q_s(b)+u)}{2pq_s(b)}-1}}{\left(\frac{n(q_s(b)+u)}{2pq_s(b)}\right)^{\frac{n(q_s(b)+u)}{2pq_s(b)}}}\right)^{\frac{pq_s(b)}{n}} \\
&\leq \exp\left(\frac{pq_s(b)}{n}\left(\frac{n(q_s(b)+u)}{2pq_s(b)} - 1 - \frac{n(q_s(b)+u)}{2pq_s(b)}\log\frac{n(q_s(b)+u)}{2pq_s(b)}\right)\right) \\
&\leq \exp\left(\frac{q_s(b)+u}{2} - \frac{pq_s(b)}{n} - \frac{q_s(b)+u}{2}\log\frac{n(q_s(b)+u)}{2pq_s(b)}\right) \\
&\leq \exp\left(-\frac{q_s(b)+u}{2}\left(\log n - \log p + \log\frac{(q_s(b)+u)}{2q_s(b)} - 1\right) - \frac{pq_s(b)}{n}\right).
\end{aligned} \tag{6}$$

Since $p = O(n^\alpha)$ for some $0 < \alpha < 1$, there exist some $N_1 \in \mathbb{N}$ and $A > 0$ such that $p \leq An^\alpha, \forall n \geq N_1$. This implies that for $n \geq N_1$, we have $\log n - \log p \geq (1-\alpha)\log n - \log A$, which, used in (6), gives

$$\mathbb{P}\left(B_s(b) \geq (q_s(b)+u)/2\right) \leq \exp\left(-\frac{q_s(b)+u}{2}\left((1-\alpha)\log n + \log\frac{(q_s(b)+u)}{2q_s(b)A} - 1\right) - \frac{pq_s(b)}{n}\right).$$

Using the fact that $q_s(b) \geq u$ for any $b \in \mathcal{B}_u$, we get

$$\mathbb{P}\left(B_s(b) \geq (q_s(b)+u)/2\right) \leq \exp\left(-u\left((1-\alpha)\log n - \log(2A) - 1\right)\right).$$

Taking $u = 2/(1-\alpha)$ and using (5), we get

$$\mathbb{P}\left(\phi(\chi_p, T) < p - \frac{2}{1-\alpha}\right) \leq \sum_{b \in \mathcal{B}_{2/(1-\alpha)}, s \in \{l,h\}} \exp\left(-2\log n + const\right) < 2n\frac{g}{n^2} = \frac{2g}{n},$$

where g is a constant and we used the fact that $|\mathcal{B}_u \times \{l,h\}| < 2n$ for any u. It follows that $\mathbb{P}(\phi(\chi_p, T) \geq p - 2/(1-\alpha)) > 1 - O(\frac{1}{n})$ as required.

Case 2. $p = cn + o(n)$ for some $0 < c < \frac{1}{2}$.

Applying (4) with $t = (q_s(b) + u)/2 - pq_s(b)/n = q_s(b)(\frac{1}{2} - \frac{p}{n}) + \frac{u}{2}$, we get

$$\begin{aligned}\mathbb{P}\left(B_s(b) \geq \frac{q_b + u}{2}\right) &\leq \exp\left(-2\left[q_s(b)\left(\frac{1}{2} - \frac{p}{n}\right) + \frac{u}{2}\right]^2 / q_s(b)\right) \\ &= \exp\left(-2\frac{q_s^2(b)(\frac{1}{2} - \frac{p}{n})^2 + 2q_s(b)(\frac{1}{2} - \frac{p}{n})\frac{u}{2} + \frac{u^2}{4}}{q_s(b)}\right) \\ &= \exp\left(-2q_s(b)\left(\frac{1}{2} - \frac{p}{n}\right)^2 - 2\left(\frac{1}{2} - \frac{p}{n}\right)u - \frac{u^2}{2q_s(b)}\right).\end{aligned} \tag{7}$$

Dropping the last term in the exponent and again using the fact that $q_s(b) \geq u$, we get

$$\begin{aligned}\mathbb{P}\left(B_s(b) \geq (q_s(b) + u)/2\right) &\leq \exp\left(-2q_s(b)\left(\frac{1}{2} - \frac{p}{n}\right)^2 - 2\left(\frac{1}{2} - \frac{p}{n}\right)u\right) \\ &\leq \exp\left(-2u\left[\left(\frac{1}{2} - \frac{p}{n}\right)^2 + \left(\frac{1}{2} - \frac{p}{n}\right)\right]\right).\end{aligned}$$

Recall that $p = cn + o(n)$. Let $c < c + \epsilon < 1/2$. For some $N_2 \in \mathbb{N}$, we will have that $p/n < c + \epsilon$ for all $n \geq N_2$, so we can take $u = C \log n$, with $C = \left[(\frac{1}{2} - (c + \epsilon))^2 + (\frac{1}{2} - (c + \epsilon))\right]^{-1}$ and use (5) to get that

$$\mathbb{P}\left(\phi(\chi_p, T) < p - C \log n\right) < 2n \exp(-2 \log n) < \frac{2}{n},$$

which gives $\mathbb{P}(\phi(\chi_p, T) \geq p - C \log n) > 1 - O(\frac{1}{n})$ as required.

Case 3. $p = \frac{1}{2}n - o(n)$

Since $\frac{p}{n} \to \frac{1}{2}$ as $n \to \infty$, the first two terms in the exponent on the right-hand side of (7) tend to 0, so the bound from the previous case is no longer useful. Based on (7), we can write

$$\mathbb{P}(B_s(b) \geq (q_s(b) + u)/2) \leq \exp\left(-\frac{u^2}{2q_s(b)}\right).$$

Knowing that $q_s(b) < n$ for any choice of b, we can take $u = C\sqrt{n \log n}$, which gives

$$\mathbb{P}\left(B_s(b) \geq (q_s(b) + C\sqrt{n \log n})/2\right) < \exp\left(-\frac{C^2 n \log n}{2n}\right) = \exp\left(-\frac{1}{2}C^2 \log n\right).$$

Setting $C = 2$ and using (5), we get

$$\mathbb{P}\left(\phi(\chi_p, T) < p - 2\sqrt{n \log n}\right) < 2n \exp(-2 \log n) = \frac{2}{n},$$

which gives us the result. □

5 Exact distribution of the transfer index on caterpillar trees

In this section, we provide exact formulae for the transfer index distribution on caterpillar trees. We shall see in the discussion section that these formulae can be used to compute p-values for the general case, under suitable assumptions (conjectures). Moreover, the combinatorial techniques used here could potentially help obtain similar results with other trees (e.g. perfectly balanced).

As a reminder, a caterpillar tree is a binary phylogenetic tree for which the induced subtree on the interior vertices forms a path graph (see Fig. 3, left). A cherry is a pair of adjacent tips on a tree. There is a single unlabeled topology for a caterpillar tree with n leaves. To identify the leaves conveniently, we label them using the natural ordering induced by the caterpillar tree topology. The tips in the two cherries have labels 1, 2, $n-1$ and n, and the other tips are labeled accordingly (Fig. 3, left). In what follows, we use T to denote the caterpillar tree labeled in that manner. This labeling/ordering is not unique, but the results are independent of the labeling options for the cherries. Since we study the distribution of the transfer index $\phi(\chi_p, T)$ where χ_p is uniformly chosen from $\mathcal{X}_p$, all bicolorations are equally probable, and our results remain identical with other labeling options. We call the tree T endowed with such labeling and coloration a bicolored oriented caterpillar tree.

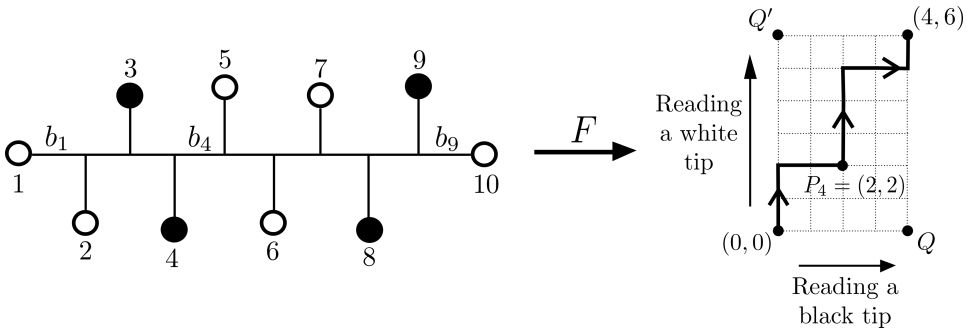


Figure 3: Left: a bicolored oriented caterpillar tree on $n = 10$ tips with $p = 4$; tips are numbered from 1 to 10 and branches on the path from tip 1 to tip 10 are denoted b_1 to b_9. Right: the associated NE lattice path from $(0, 0)$ to $(4, 6)$; the point $P_4 = (2, 2)$ on the path corresponds to branch b_4 on the tree. Center: the function F that associates the tree on the left with the path on the right.

5.1 Correspondence between bicolored caterpillar trees and NE lattice paths

An *NE lattice path* is a path in $\mathbb{Z}^2$ where the only steps allowed are $(0,1)$ (a step towards the east) and $(1,0)$ (a step towards the north). From now on, we call them lattice paths for short. Let $\mathcal{P}(p, n-p)$ denote the $p \times (n-p)$ NE lattice, that is the set of all lattice paths from the origin $(0,0)$ to the destination $(p, n-p)$. A path in $\mathcal{P}(p, n-p)$ can be encoded in a single vector of length n indicating the sequence of steps of the path, which is an element on $\{N, E\}^n$ with a total number of east steps equal to p and a total number of north steps equal to $n-p$. On the other hand, the set $\mathcal{X}_p$ of bicolorations with p black tips is a subset of $\{\mathbf{B}, \mathbf{W}\}^n$. We define the function $F : \mathcal{X}_p \to \mathcal{P}(p, n-p)$ that associates a lattice path to a bicolored tree by scanning the tips on the tree from 1 to n as follows: whenever we read a white leaf, we move towards the north; and whenever we read a black leaf, we move towards the east. Consequently, a bicolored oriented caterpillar tree on n tips (with p black tips) corresponds to a unique path in $\mathcal{P}(p, n-p)$ and vice versa, as represented in Fig. 3. This result can be summarized as follows.

Lemma 10. *The function F is a bijection from $\mathcal{X}_p$ to $\mathcal{P}(p, n-p)$.*

Let us denote the lower right corner in $\mathcal{P}(p, n-p)$ by $Q = (p, 0)$ and the upper left corner by $Q' = (0, n-p)$. Observe that the two extreme paths going through Q and Q' correspond to the only bicolored oriented caterpillar trees T with transfer index $\phi(\chi_p, T) = 0$: all black leaves cluster on one side and all white leaves on the other side. These two extreme paths can be identified with the reference bipartition β (Fig. 4a, left, in green). Moreover, we are able to retrieve the transfer index for any bicolored oriented caterpillar tree from the associated lattice path, as we demonstrate in the following proposition. Use $M(A, B)$ to denote the Manhattan distance between any two lattice points $A, B \in \mathbb{Z}^2$, and by $M(\gamma, B) = \min_{A \in \gamma} M(A, B)$ the Manhattan distance between any lattice path γ and a lattice point B.

Proposition 11. *Consider an oriented caterpillar tree T, a bicoloration $\chi_p \in \mathcal{X}_p$ of its tips, and the corresponding path $\gamma \in \mathcal{P}(p, n-p)$. We have that*

$$\phi(\chi_p, T) = \min\left(M\left(\gamma, Q\right), M\left(\gamma, Q'\right), p-1\right).$$

Proof. Consider an oriented caterpillar tree T, a bicoloration χ_p, and the corresponding lattice path γ from $(0,0)$ to $(p, n-p)$. Let us denote the $n-1$ consecutive internal lattice points in γ by $P_1 = (x_1, y_1), \ldots, P_{n-1} = (x_{n-1}, y_{n-1})$. Also, use b_i to denote the internal branch in T between tips i and $i+1$, for $2 \leq i \leq n-2$. Lastly, let b_1 and b_{n-1} be the pendant branches of tips 1 and n respectively (Fig. 3, left).

In the same manner that F associates tips in the bicolored tree with steps in γ, this function can be extended naturally so it establishes a one-to-one mapping from a set of branches in T to the interior lattice points in γ. More precisely, if we extend F to the set of branches in T, it holds that $F(b_i) = P_i$, for all $1 \leq i \leq n-1$. The transfer distance between χ_p and an internal branch b_i is the number of tips to be transferred from one side of b_i to the other side that results in the bipartition χ_p. Fix $1 \leq i \leq n-1$ and consider the left side of b_i. Use $W(b_i)$ and $B(b_i)$ to denote respectively the number of black and white tips in this left side. By construction, we have that $W(b_i) = y_i$ and $B(b_i) = x_i$, which, together with (2), leads to the following identity

$$\delta(\chi_p, b_i) = \min(x_i + n - p - y_i, p - x_i + y_i).$$

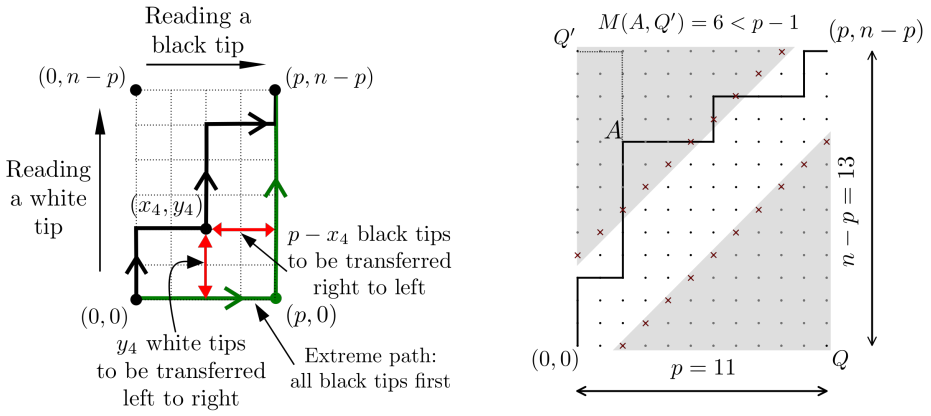
On the other hand, if we look at the Manhattan distance between the corresponding lattice point P_i and the corners Q and Q', we have that $M(P_i, Q) = p - x_i + y_i$ and $M(P_i, Q') = x_i + n - p - y_i$ (see Fig. 4a). The same argument applies to any branch b_i for $1 \leq i \leq n-1$. Hence, the minimum of these distances taken over all lattice points P_i in the path γ corresponds to the minimum of the transfer distance obtained by any internal branch in T, or by the leaves 1 and n.

Finally, all branches on the caterpillar tree that are not on the path from leaf 1 to leaf n are pendant branches. The minimum over all the pendant branches is equal to $p-1$, obtained on any black leaf, so the transfer index is at least $p-1$, as stated in the proposition. Also notice that, in the case of a bicolored cherry, the choice of the labels ($1, 2$ or $n-1, n$) has no influence on the result since the distance from any of these pendant branches to the reference bipartition is at least $p-1$. Since we have covered the distance obtained on any branch on the tree, we achieve the desired result. □

5.2 Counting bicolorations through lattice paths: the transfer index distribution

Lattice paths under certain restrictions appear in various problems in probability and statistics, such as the classical ballot problem (for instance, see [10]), which leads to counting lattice paths that do not touch the diagonal $y = x$. Here, we are interested in a slightly different problem, but closely related to the ballot problem in the sense that we count NE lattice paths that are not allowed to touch certain boundaries.

More precisely, for fixed n and p, consider $2 \leq l \leq p+1$ and let $\mathscr{L}(n, p, l)$ denote the subset of paths in $\mathcal{P}(p, n-p)$ that do not touch $y = x - l$ or $y = x + (n - 2p + l)$. Set $L(n, p, l) := |\mathscr{L}(n, p, l)|$. The following result is from Mohanty [20]. Here, we will give a sketch of the proof that is slightly different from the one in [20], since it will be useful for understanding the



(a) (b)

Figure 4: (a) Green: the extreme lattice path corresponding to the $\{\mathbf{B}\}$ first then $\{\mathbf{W}\}$ bicoloration. The branch b corresponding to the point $(p, 0)$ in the path, results in $\delta(\chi_p, b) = 0$. On the lattice path in bold lines, the point (x_4, y_4) corresponds to the branch b_4 of the corresponding tree T with $M((x_4, y_4), (p, 0)) = p - 2 + 2 = 4$ (i.e. the transfer distance of b_4 as shown in Fig. 3). (b) Paths avoiding the shaded areas have a Manhattan distance to the corners Q and Q' larger than $p - 1$, and thus the corresponding bicolorations yield a transfer index $= p - 1$. When a path enters a shaded area, its Manhattan distance to the corners is less than $p-1$, e.g. point A is at distance 6 of Q' and thus the corresponding path and bicoloration have a transfer index $= 6$.

upcoming results. We use $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ to denote respectively floor and ceiling functions, and $\mathbb{1}(\cdot)$ for the indicator function.

Lemma 12 ([20]). *Let $c = n - 2p + 2l$, then*

$$L(n,p,l) = \sum_{k=\left\lfloor \frac{p-l-n}{c} \right\rfloor}^{\left\lceil \frac{p}{c} \right\rceil} \left[\binom{n}{p-kc} - \binom{n}{p-l-kc} \right].$$

Proof. The proof is based on the well-known André's reflection method [1]. First, notice that a path is uniquely defined by the p east steps it makes (or equivalently the $n - p$ north steps), which entails that

$$|\mathcal{P}(p, n-p)| = \binom{n}{p}.$$

Let us count now the paths in $\mathcal{P}(p, n-p)$ that touch the line $y = x - l$. If a path touching the line $y = x - l$ is reflected from the moment it first touches

this line, by switching north steps into east steps and vice versa, we obtain a lattice path ending at $(n-p+l, p-l)$. In fact, this reflection yields a bijection between the set of paths in $\mathcal{P}(p, n-p)$ touching the line $y = x - l$ and the set $\mathcal{P}(n-p+l, p-l)$, both having $\binom{n}{p-l}$ elements.

Likewise, paths in $\mathcal{P}(p, n-p)$ that touch the line $y = x + (n-2p+l)$ can be transformed bijectively into paths in the set $\mathcal{P}(p-l, n-p+l)$. Then, by applying the *inclusion-exclusion principle* [13], we have the following identity,

$$L(n,p,l) = \binom{n}{p} - 2\binom{n}{p-l} + |\mathcal{L}'(n,p,l)|, \tag{8}$$

where $\mathcal{L}'(n,p,l)$ is the set of paths that touch both lines $y = x - l$ and $y = x+(n-2p+l)$. We can then apply the reflection principle repeatedly and the usual inclusion-exclusion principle to account for paths touching both lines multiple times, leading to the above formula after some simplifications. See [20] for further details. □

We can now establish the main theorem in this section.

Theorem 13. *Let T be an oriented caterpillar tree and χ_p a bicoloration uniformly chosen from the set $\mathcal{X}_p$. We have for any $2 \leq l \leq p+1$ that*

$$\mathbb{P}\left(\phi\left(\chi_p, T\right) = p - l + 1\right) = \frac{L(n,p,l) - L(n,p,l-1)\mathbb{1}(l>2)}{\binom{n}{p}} \tag{9}$$

and

$$\mathbb{E}\left[\phi\left(\chi_p, T\right)\right] = \frac{1}{\binom{n}{p}} \sum_{l=2}^{p} L(n,p,l). \tag{10}$$

Proof. It is quite straightforward from the definition of $\mathscr{L}(n,p,l)$ and Prop. 11 that for any $2 \leq l \leq p+1$, the paths avoiding lines $y = x - l$ and $y = x + (n-2p+l)$ are exactly those that remain at a distance from the corners Q and Q' greater or equal to $p-l+1$, as it is shown in Fig. 4b, for $l = 2$. On the other hand, the paths that do touch these lines give a distance strictly smaller than $p - l + 1$. Hence, for $2 \leq l \leq p+1$, a path in the set $\mathscr{L}(n,p,l) \setminus \mathscr{L}(n,p,l-1)$ (with the convention $\mathscr{L}(n,p,1) = \emptyset$) gives a distance exactly equal to $p-l+1$. Since the total number of paths in $\mathcal{P}(p, n-p)$ is $\binom{n}{p}$, the identity (9) holds. Then, the expectation of the transfer index can be expressed as follows

$$E\left[\phi\left(\chi_p, T\right)\right] = \frac{1}{\binom{n}{p}} \sum_{l=2}^{p+1} \left[L(n,p,l) - L(n,p,l-1)\mathbb{1}(l>2)\right](p-l+1),$$

which can easily be simplified to obtain (10). □

6 Discussion

The results we obtained in Sections 3 and 4 allow us to characterize the asymptotic behavior of the transfer support when n tends towards ∞, for various growth rates of p, up to $p = \lfloor n/2 \rfloor$ and for any tree topology. However, the bounds we obtained in Section 4 are not sufficiently tight to justifiy what we observe from our simulations in Fig. 1. If we think of applications, these bounds might not be sufficient to give good estimates for the p-values of the TS distribution. We propose two conjectures that allow us to use the exact results obtained for the caterpillar tree as a proxy for the statistical significance of TS on the null models.

The first conjecture concerns the extreme case $p = \lfloor n/2 \rfloor$. Based on simulation results (Fig. 1 and [18]) and the proofs in section 4, we believe that for any tree topology, the expected value of TS attains its maximum over p at $\lfloor n/2 \rfloor$. The second conjecture (based on Fig. 1 and not shown experiments) refers to the stochastic dominance of TS for the caterpillar tree over TS for any other tree topology at $p = \lfloor n/2 \rfloor$.

Assuming that these conjectures (or similar ones) hold, we can bound the p-values for any p and any topology by the caterpillar case at $\lfloor n/2 \rfloor$, for which we have an explicit formula. For instance, consider a tree on 100 tips and a symetric reference bipartition, that is $p = 50$. The probability of this branch being supported at 50% on a caterpillar tree under the null model (the absence of phylogenetic signal) is $\sim 10^{-6}$. For a tree on 20 tips and $p = 10$, the same probability is $\sim 5\%$. These values support the idea discussed in [18] that standard levels of branch support using TBE (say $> 70\%$, following Hillis and Bull [16]) cannot be observed by chance, and reveal a strong phylogenetic signal in the data, even with small trees.

For trees that are not caterpillars, deriving the distribution of the transfer index under random bicolorations appears to be challenging. It would be relevant for both theoretical and applicative reasons to characterize this distribution for a random model such as Yule or PDA, which are the most commonly used in phylogenetics [14].

7 Appendix: asymptotics of the transfer index on the caterpillar tree

We obtained an additional result describing the asymptotic behavior of the transfer index when n gets large, in the case of caterpillar trees. Similarly to our results in Sections 3 and 4, the following proposition implies that TS tends to 0 when $n \to \infty$ for random bicolorations. However, the speed of convergence implied by the proposition below improves the one obtained in previous sections. In particular, for $p = \lfloor n/2 \rfloor$, Theorem 8 implies that the expectation of TS grows at most as $\sqrt{\log n}/\sqrt{n}$, whereas for the caterpillar

tree it is $1/\sqrt{n}$ as we demonstrate below.

Proposition 14. *Let T be an oriented caterpillar tree and χ_p a bicoloration uniformly chosen from the set $\mathcal{X}_p$, for $2 \leq p \leq \lfloor n/2 \rfloor$. The expected transfer support goes to 0 uniformly on p, moreover*

$$\max_{p \leq n/2} \left(1 - \frac{\mathbb{E}\left[\phi\left(\chi_p, T\right)\right]}{p-1}\right) = O\left(\frac{1}{\sqrt{n}}\right), \text{ when } n \to \infty. \tag{11}$$

Proof. For any n, p, l with $p \leq n/2$ and $2 \leq l \leq p+1$ we have from (8) that

$$L(n,p,l) \geq \binom{n}{p} - 2\binom{n}{p-l}.$$

This inequality, together with (10), leads to

$$\mathbb{E}\left[\phi\left(\chi_p, T\right)\right] \geq p - 1 - \frac{2}{\binom{n}{p}} \sum_{l=2}^{p} \binom{n}{p-l},$$

or equivalently,

$$\mathbb{E}\left[\mathrm{TS}\left(\chi_p, T\right)\right] = 1 - \frac{\mathbb{E}\left[\phi\left(\chi_p, T\right)\right]}{p-1} \leq \frac{2}{(p-1)\binom{n}{p}} \sum_{l=2}^{p} \binom{n}{p-l} =: G(p).$$

Let us now prove that $G(p)$ is an increasing function for $2 \leq p \leq \lfloor n/2 \rfloor$,

$$\begin{aligned} G(p+1) - G(p) &= \frac{2}{(p)\binom{n}{p+1}} \sum_{k=0}^{p-1} \binom{n}{k} - \frac{2}{(p-1)\binom{n}{p}} \sum_{k=0}^{p-2} \binom{n}{k} \\ &= 2\sum_{k=1}^{p-1} \left[\frac{\binom{n}{k}}{p\binom{n}{p+1}} - \frac{\binom{n}{k-1}}{(p-1)\binom{n}{p}}\right] + \frac{2}{p\binom{n}{p+1}} \\ &= 2\sum_{k=1}^{p-1} \frac{(p+1)!(n-p-1)!}{(k-1)!(n-k+1)!} \left[\frac{n-k+1}{pk} - \frac{n-p}{p^2-1}\right] + \frac{2}{p\binom{n}{p+1}}, \end{aligned}$$

where the terms between brackets in the sum can be expanded as follows

$$\frac{n-k+1}{pk} - \frac{n-p}{p^2-1} = \frac{n(p^2-kp-1)+p^2+k-1}{pk(p^2-1)}.$$

It is not hard to see that $p^2 - kp - 1 \geq 0$ for all values of k and p such that $1 \leq k \leq p-1$, which implies that the previous fraction is always positive. We can then conclude that $G(p+1) - G(p) \geq 0$, so the maximum value of G is obtained when $p = \lfloor n/2 \rfloor$.

For the sake of simplicity, we suppose from now that n is even, but an equivalent result is obtained for odd n without difficulty. For $p = n/2$,

we can use the identity $\sum_{k=0}^{n}\binom{n}{k} = 2^n$ and the symmetry of the binomial coefficients $\binom{n}{k} = \binom{n}{n-k}$ for all $0 \leq k \leq n$, to simplify the sum in the function G as follows,

$$\sum_{l=2}^{n/2}\binom{n}{n/2-l} = \sum_{k=0}^{n/2-2}\binom{n}{k} = \frac{2^n - 2\binom{n}{n/2-1} - \binom{n}{\lfloor n/2 \rfloor}}{2}.$$

So, for all $2 \leq p \leq n/2$, we have that

$$0 \leq \mathbb{E}\left[\mathrm{TS}\left(\chi_p, T\right)\right] \leq G\left(n/2\right) = \frac{2^n - 2\binom{n}{n/2-1} - \binom{n}{n/2}}{\left(n/2-1\right)\binom{n}{n/2}}.$$

Let us now look at the asymptotic behavior of this upper bound. The well-known Stirling formula yields the following approximation for central binomial coefficients $\binom{n}{n/2} \sim \frac{2^{n+1/2}}{\sqrt{\pi n}}$. Hence

$$\frac{2^n - 2\binom{n}{n/2-1} - \binom{n}{n/2}}{\left(n/2-1\right)\binom{n}{n/2}} \sim \frac{\sqrt{2\pi n}}{n-2} - \frac{4n}{(n-2)(n+2)} - \frac{2}{n-2} \sim \frac{\sqrt{2\pi}}{\sqrt{n}},$$

which tends towards 0 when $n \to \infty$, allowing conclusion of the proof. □

Acknowledgements

This work was supported by the EU-H2020 Virogenesis project (grant number 634650 – JT, OG), by the INCEPTION project (PIA/ANR-16-CONV-0005 – MDF, FL, OG).

References

[1] D. André. Solution directe du problème résolu par M. Bertrand. *C.R. Acad. Sci., Paris*, 105:436–437, 1887.

[2] A. Boc, H. Philippe, and V. Makarenkov. Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Systematic Biology*, 59(2):195–211, 2010.

[3] D. Bogdanowicz and K. Giaro. Matching split distance for unrooted binary phylogenetic trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(1):150–160, Jan. 2012.

[4] I. Charon, L. Denœud, A. Guénoche, and O. Hudry. Maximum transfer distance between partitions. *Journal of Classification*, 23(1):103–121, June 2006.

[5] W. H. Day. The complexity of computing metric distances between partitions. *Mathematical Social Sciences*, 1(3):269 – 287, 1981.

[6] W. H. E. Day. Optimal algorithms for comparing trees with labeled leaves. *Journal of Classification*, 2(1):7–28, Dec. 1985.

[7] L. Denœud. Transfer distance between partitions. *Advances in Data Analysis and Classification*, 2(3):279–294, Dec. 2008.

[8] D. P. Dubhashi and A. Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[9] J. S. Farris. Methods for computing wagner trees. *Systematic Zoology*, 19(1):83–92, 1970.

[10] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, Jan. 1968.

[11] J. Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39:783–791, 1985.

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

[13] M. Fréchet. *Les probabilités associées à un système d'événements compatibles et dépendants*. Actualités scientifiques et industrielles. Hermann & Cie, 1940 et 1943.

[14] O. Gascuel. *Mathematics of Evolution and Phylogeny*. Oxford University Press, Inc., New York, NY, USA, 2007.

[15] J. A. Hartigan. Minimum mutation fits to a given tree. *Biometrics*, 29(1):53–65, 1973.

[16] D. M. Hillis and J. J. Bull. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, 42(2):182–192, 1993.

[17] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[18] F. Lemoine, J.-B. Domelevo Entfellner, E. Wilkinson, D. Correia, M. Dávila Felipe, T. De Oliveira, and O. Gascuel. Renewing Felsenstein's Phylogenetic Bootstrap in the Era of Big Data. *Nature*, 556(7702):452–456, Apr. 2018.

[19] Y. Lin, V. Rajan, and B. M. E. Moret. A metric for phylogenetic trees based on matching. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 9(4):1014–1022, 2012.

[20] G. Mohanty. *Lattice path counting and applications*. Probability and mathematical statistics. Academic Press, 1979.

[21] D. Robinson and L. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131 – 147, 1981.

[22] S. Régnier. Sur quelques aspects mathématiques des problèmes de classification automatique. *ICC Bulletin*, 4:175–191, 1965.

[23] C. Semple and M. Steel. *Phylogenetics*. Oxford University Press, Oxford New York, 2003.

[24] M. Steel. *Phylogeny: Discrete and Random Processes in Evolution*. CBMS-NSF Regional Conference Series on Mathematics. Society for Industrial and Applied Mathematics, 2016.

[25] M. Steel and D. Penny. Maximum parsimony and the phylogenetic information in multi-state characters. In V. Albert, editor, *Parsimony, phylogeny and genomics*, chapter 9, pages 163–178. Oxford University Press, 2005.