ON SYMMETRIES IN PHYLOGENETIC TREES

ÉRIC FUSY*

Abstract. Billey et al. [arXiv:1507.04976] have recently discovered a surprisingly simple formula for the number $a_n(\sigma)$ of leaf-labelled rooted non-embedded binary trees (also known as phylogenetic trees) with $n \geq 1$ leaves, fixed (for the relabelling action) by a given permutation $\sigma \in \mathfrak{S}_n$. Denoting by $\lambda \vdash n$ the integer partition giving the sizes of the cycles of σ in non-increasing order, they show by a guessing/checking approach that if λ is a binary partition (it is known that $a_n(\sigma) = 0$ otherwise), then

$$a_n(\sigma) = \prod_{i=2}^{\ell(\lambda)} (2(\lambda_i + \cdots + \lambda_{\ell(\lambda)}) - 1),$$

and they derive from it a formula and random generation procedure for tanglegrams (and more generally for tangled chains). Our main result is a combinatorial proof of the formula for $a_n(\sigma)$, which yields a simplification of the random sampler for tangled chains.

1. Introduction

For A a finite set of cardinality $n \geq 1$, we denote by $\mathcal{B}[A]$ the set of rooted binary trees that are non-embedded (i.e., the order of the two children of each node does not matter) and have n leaves with distinct labels from A. Such trees are known as *phylogenetic trees*, where typically A is the set of represented species. Note that such a tree has $n-1$ nodes and $2n-1$ edges (we take here the convention of having an additional root-edge above the root-node, connected to a 'fake-vertex' that does not count as a node, see Figure 1).

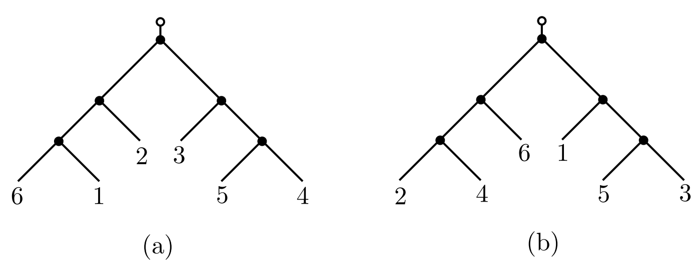

Figure 1. (a) A phylogenetic tree γ with label-set [1..6]. (b) The tree $\gamma' = \sigma \cdot \gamma$, with $\sigma = (1,4,3)(5)(2,6)$. Since $\gamma' \neq \gamma$, γ is not fixed by σ (on the other hand γ is fixed by $(2,3)(1,4,6,5)$).

The group $\mathfrak{S}(A)$ of permutations of A acts on $\mathcal{B}[A]$: for $\gamma \in \mathcal{B}[A]$ and $\sigma \in \mathfrak{S}(A)$, $\sigma \cdot \gamma$ is obtained from γ after replacing the label i of every leaf by $\sigma(i)$, see

*LIX, École Polytechnique, Palaiseau, France, fusy@lix.polytechnique.fr. Partly supported by the ANR grant "Cartaplus" 12-JS02-001-01 and the ANR grant "EGOS" 12-JS02-002-01.

Figure 1(b). We denote by $\mathcal{B}_\sigma[A]$ the set of trees fixed by the action of σ, i.e., $\mathcal{B}_\sigma[A] := \{\gamma \in \mathcal{B}[A] \text{ such that } \sigma \cdot \gamma = \gamma\}$. We also define $\mathcal{E}_\sigma[A]$ (resp. $\mathcal{E}[A]$) as the set of pairs (γ, e) where $\gamma \in \mathcal{B}_\sigma[A]$ (resp. $\gamma \in \mathcal{B}[A]$) and e is an edge of γ (among the $2n-1$ edges). Define the *cycle-type* of σ as the integer partition $\lambda \vdash n$ giving the sizes of the cycles of σ (in non-increasing order). For $\lambda \vdash n$ an integer partition, the cardinality of $\mathcal{B}_\sigma[A]$ is the same for all permutations σ with cycle-type λ, and this common cardinality is denoted by r_λ. It is known (e.g. using cycle index sums [1, 3]) that $r_\lambda = 0$ unless λ is a binary partition (i.e., an integer partition whose parts are powers of 2). Billey et al. [2] have recently found the following remarkable formula, valid for any binary partition λ:

$$r_\lambda = \prod_{i=2}^{\ell(\lambda)} (2(\lambda_i + \cdots + \lambda_{\ell(\lambda)}) - 1). \tag{1}$$

They prove the formula by a guessing/checking approach. Our main result here is a combinatorial proof of (1), which yields a simplification (see Section 3) of the random sampler for tanglegrams (and more generally tangled chains) given in [2].

Theorem 1. *For A a finite set and σ a permutation on A whose cycle-type is a binary partition:*

- *If σ has one cycle, then $|\mathcal{B}_\sigma[A]| = 1$.*
- *If σ has more than one cycle, let c be a largest cycle of σ; denote by A' the set A without the elements of c, and denote by σ' the permutation σ restricted to A'. Then we have the combinatorial isomorphism*

$$\mathcal{B}_\sigma[A] \simeq \mathcal{E}_{\sigma'}[A']. \tag{2}$$

As we will see, the isomorphism (2) can be seen as an adaptation of Rémy's method [7] to the setting of (non-embedded rooted) binary trees fixed by a given permutation. Note that Theorem 1 implies that the coefficients r_λ satisfy $r_\lambda = 1$ if λ is a binary partition with one part and $r_\lambda = (2|\lambda\backslash\lambda_1| - 1) \cdot r_{\lambda\backslash\lambda_1}$ if λ is a binary partition with more than one part, from which we recover (1).

2. Proof of Theorem 1

2.1. **Case where the permutation σ has one cycle.** The fact that $|\mathcal{B}_\sigma[A]| = 1$ if σ has one cycle of size 2^k (for some $k \geq 0$) is well known from the structure of automorphisms in trees [6], for the sake of completeness we give a short justification. Since the case $k = 0$ is trivial we can assume that $k \geq 1$. Let c_1, c_2 be the two cycles of σ^2 (each of size 2^{k-1}), with the convention that c_1 contains the minimal element of A; denote by A_1, A_2 the induced bi-partition of A, and by $\sigma_1 = c_1$ (resp. $\sigma_2 = c_2$) the permutation σ^2 restricted to A_1 (resp. A_2). For $\gamma \in \mathcal{B}_\sigma[A]$ let γ_1, γ_2 be the two subtrees at the root-node of γ, such that the minimal element of A is in γ_1. Then clearly $\gamma_1 \in \mathcal{B}_{\sigma_1}[A_1]$ and $\gamma_2 \in \mathcal{B}_{\sigma_2}[A_2]$, and conversely for $\gamma_1 \in \mathcal{B}_{\sigma_1}[A_1]$ and $\gamma_2 \in \mathcal{B}_{\sigma_2}[A_2]$ the tree γ with (γ_1, γ_2) as subtrees at the root-node is in $\mathcal{B}_\sigma[A]$. Hence

$$\mathcal{B}_\sigma[A] \simeq \mathcal{B}_{\sigma_1}[A_1] \times \mathcal{B}_{\sigma_2}[A_2], \tag{3}$$

which implies $|\mathcal{B}_\sigma[A]| = 1$ by induction on k (note that, also by induction on k, the underlying unlabelled tree is the complete binary tree of height k).

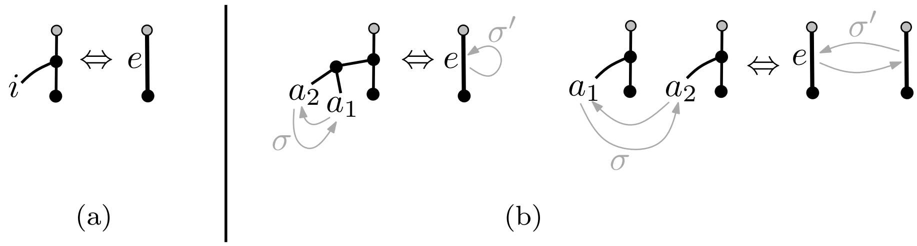

FIGURE 2. (a) Rémy's leaf-removal operation. (b) The two cases for removing a 2-cycle of leaves (depending whether the two leaves have the same parent or not). The vertices depicted gray are allowed to be the fake vertex above the root-node.

2.2. **Case where the permutation σ has more than one cycle.** Let $k \geq 0$ be the integer such that the largest cycle of σ has size 2^k. A first useful remark is that σ induces a permutation of the edges (resp. of the nodes) of γ, and each σ-cycle of edges (resp. of nodes) has size 2^i for some $i \in [0..k]$. We present the proof of (2) progressively, treating first the case $k = 0$, then $k = 1$, then general k.

Case $k = 0$. This case corresponds to σ being the identity, so that $\mathcal{B}_\sigma[A] \simeq \mathcal{B}[A]$, hence we just have to justify that $\mathcal{B}[A] \simeq \mathcal{E}[A \backslash \{i\}]$ for each fixed $i \in A$. This is easy to see using Rémy's argument [7] [1], used here in the non-embedded leaf-labelled context: every $\gamma \in \mathcal{B}[A]$ is uniquely obtained from some $(\gamma', e) \in \mathcal{E}[A \backslash \{i\}]$ upon inserting a new pending edge from the middle of e to a new leaf that is given label i, see Figure 2(a).

Case $k = 1$. Let $c = (a_1, a_2)$ be the selected cycle of σ, with $a_1 < a_2$. Two cases can arise (in each case we obtain from γ a pair (γ', e) with $\gamma' \in \mathcal{B}_{\sigma'}[A']$ and e an edge of γ'):

- if a_1 and a_2 have the same parent v, we obtain a reduced tree $\gamma' \in \mathcal{B}_{\sigma'}[A']$ by erasing the 3 edges incident to v (and the endpoints of these edges, which are a_1, a_2, v and the parent of v), and we mark the edge e of γ' whose middle was the parent of v, see the first case of Figure 2(b)
- if a_1 and a_2 have distinct parents, we can apply the operation of Figure 2(a) to each of a_1 and a_2, which yields a reduced tree $\gamma' \in \mathcal{B}_{\sigma'}[A']$. We then mark the edge e of γ' whose middle was the parent of a_1, see the second case of Figure 2(b).

Conversely, starting from $(\gamma', e) \in \mathcal{E}[A']$, the σ'-cycle of edges that contains e has either size 1 or 2:

- if it has size 1 (i.e., e is fixed by σ'), we insert a pending edge from the middle of e and leading to "cherry" with labels (a_1, a_2),
- if it has size 2, let $e' = \sigma'(e)$; then we attach at the middle of e (resp. e') a new pending edge leading to a new leaf of label a_1 (resp. a_2).

The general case $k \geq 0$. Recall that the marked cycle of σ is denoted by c. A node or leaf of the tree is generically called a *vertex* of the tree. We define a *c-vertex* as a vertex v of γ such that:

[1] A similar argument in the context of triangulations of a polygon dates back to Rodrigues [8].

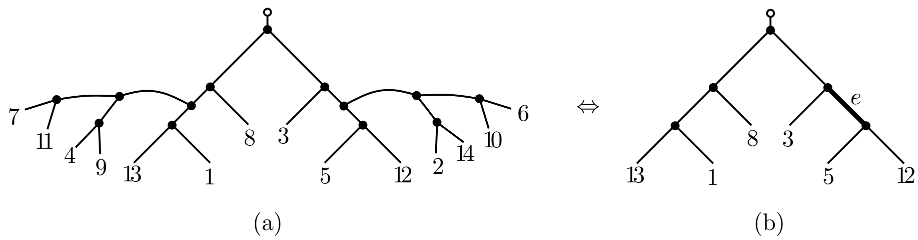

FIGURE 3. (a) a tree in $\mathcal{B}_\sigma[A]$, for $A = [1..14]$ and $\sigma = (3,8)(1,5,13,12)(2,7,10,4,14,11,6,9)$. (b) The corresponding (when selecting the cycle c of size 8 in σ) pair $(\gamma', e) \in \mathcal{E}_{\sigma'}[A']$, where $A' = A\backslash c$ and $\sigma' = (3,8)(1,5,13,12)$ (restriction of σ to A').

- if v is a leaf then $v \in c$,
- if v is a node then all leaves that are descendant of v are in c.

A c-vertex is called *maximal* if it is not the descendant of any other c-vertex; define a *c-tree* as a subtree formed by a maximal c-vertex v and its hanging subtree (if v is a leaf then the corresponding c-tree is reduced to v). Note that the maximal c-vertices are permuted by σ. Moreover since the leaves of c are permuted cyclically, the maximal c-vertices actually have to form a σ-cycle of vertices, of size 2^i for some $i \leq k$; and in each c-tree, σ^{2^i} permutes the 2^{k-i} leaves of the c-tree cyclically. Let ℓ be the leaf of minimal label in c, and let w be the maximal c-vertex such that the c-tree at w contains ℓ. We obtain a reduced tree $\gamma' \in \mathcal{B}_{\sigma'}[A']$ by erasing all c-trees and erasing the parent-edges and parent-vertices of all maximal c-vertices; and then we mark the edge e of γ' whose middle was the parent of w, see Figure 3.

Conversely, starting from $(\gamma', e) \in \mathcal{E}_{\sigma'}[A']$, let $i \in [0..k]$ be such that the σ'-cycle of edges that contains e has cardinality 2^i; write this cycle as $e_0, \ldots, e_{2^i-1}$, with $e_0 = e$. Starting from the element of c of minimal label, let $(s_0, \ldots, s_{2^i-1})$ be the 2^i (successive) first elements of c. And for $r \in [0..2^i-1]$ let c_r be the cycle of σ^{2^i} that contains s_r, and let A_r be the set of elements in c_r (note that $A_0, \ldots, A_{2^i-1}$ each have size 2^{k-i} and partition the set of elements in c). Let T_r be the unique (by Section 2.1) tree in $\mathcal{B}[A_r]$ fixed by the cyclic permutation c_r. We obtain a tree $\gamma \in \mathcal{B}_\sigma[A]$ as follows: for each $r \in [0..2^i-1]$ we create a new edge that connects the middle of e_r to a new copy of T_r.

To conclude we have described a mapping from $\mathcal{B}_\sigma[A]$ to $\mathcal{E}_{\sigma'}[A']$ and a mapping from $\mathcal{E}_{\sigma'}[A']$ to $\mathcal{B}_\sigma[A]$ that are readily seen to be inverse of each other, therefore $\mathcal{B}_\sigma[A] \simeq \mathcal{E}_{\sigma'}[A']$.

3. Application to the random generation of tangled chains

For $n \geq 1$, denote by $\mathbf{n}$ the set $\{1, \ldots, n\}$. A *tanglegram* of size n is an orbit of $\mathcal{B}[\mathbf{n}] \times \mathcal{B}[\mathbf{n}]$ under the relabelling action of $\mathfrak{S}_n$ (see Figure 4 for an example). More generally, for $k \geq 1$, a *tangled chain* of length k and size n is an orbit of $\mathcal{B}[\mathbf{n}]^k$ under the relabelling action of $\mathfrak{S}_n$, see [5, 2, 3]. Let $\mathcal{T}_n^{(k)}$ be the set of tangled chains of length k and size n, and let $t_n^{(k)}$ be the cardinality of $\mathcal{T}_n^{(k)}$. Then it follows from Burnside's lemma (see [2] for a proof using double cosets and [3] for a proof using

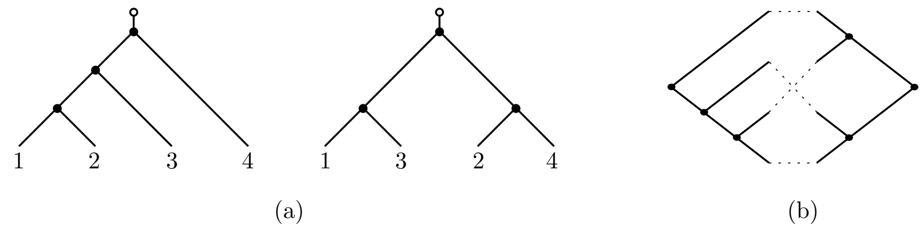

FIGURE 4. (a) A pair of (rooted non-embedded leaf-labelled) binary trees. (b) the corresponding (unlabelled) tanglegram.

the formalism of species) that

$$t_n^{(k)} = \frac{1}{n!}\sum_{\sigma\in\mathfrak{S}_n} |\mathcal{B}_\sigma[\mathbf{n}]|^k = \sum_{\lambda\vdash n} \frac{{r_\lambda}^k}{z_\lambda}, \tag{4}$$

where $z_\lambda = 1^{m_1}m_1!\cdots r^{m_r}m_r!$ if λ has m_1 parts of size 1,...,m_r parts of size r (recall that $n!/z_\lambda$ is the number of permutations with cycle-type λ). At the level of combinatorial classes, Burnside's lemma gives

$$\mathfrak{S}_n \times \mathcal{T}_n^{(k)} \simeq \sum_{\sigma\in\mathfrak{S}_n} \mathcal{B}_\sigma[\mathbf{n}]^k,$$

and thus the following procedure is a uniform random sampler for $\mathcal{T}_n^{(k)}$ (see [2] for details):

(1) Choose a random binary partition $\lambda \vdash n$ under the distribution

$$P(\lambda) = \frac{{r_\lambda}^k/z_\lambda}{S_n},$$

where $S_n = \sum_{\lambda\vdash n} {r_\lambda}^k/z_\lambda$ $(= t_n^{(k)})$.

(2) Let σ be a permutation with cycle-type λ. For each $r \in [1..k]$ draw (independently) a tree $T_r \in \mathcal{B}_\sigma[\mathbf{n}]$ uniformly at random.

(3) Return the tangled chain corresponding to $(T_1, \ldots, T_k)$.

A recursive procedure (using (1)) is given in [2] to sample uniformly at random from $\mathcal{B}_\sigma[\mathbf{n}]$. From Theorem 1 we obtain a simpler random sampler for $\mathcal{B}_\sigma[\mathbf{n}]$. We order the cycles of σ as $c_1, \ldots, c_{\ell(\lambda)}$ such that the cycle-sizes are in non-decreasing order. Then, with A_1 the set of labels in c_1, we start from the unique tree (by Section 2.1) in $\mathcal{B}_{c_1}[A_1]$ (where c_1 is to be seen as a cyclic permutation on A_1). Then, for i from 2 to $\ell(\lambda)$ we mark an edge chosen uniformly at random from the already obtained tree, and then we insert the leaves that have labels in c_i using the isomorphism (2).

The complexity of the sampler for $\mathcal{B}_\sigma[\mathbf{n}]$ is clearly linear in n and needs no precomputation of coefficients. However step (1) of the random generator requires a table of $p(n)$ coefficients, where $p(n)$ is the number of binary partitions of n, which is slightly superpolynomial [4], $p(n) = n^{\Theta(\log(n))}$. It is however possible to do step (1) in polynomial time. For this, we consider, for $i \geq 0$ and $n, j \geq 1$ the coefficient $S_n^{(i,j)}$ defined as the sum of ${r_\lambda}^k/z_\lambda$ over all binary partitions of n where the largest part is 2^i and has multiplicity j; note that $S_n^{(i,j)} = 0$ unless $j\cdot 2^i \leq n$, we denote by E_n the set of such pairs (i,j). Since $r_\lambda = 1$ and $z_\lambda = (|\lambda|-1)!$ if λ has one part, we have the initial condition $S_n^{(i,j)} = 1/(n-1)!$ for $j = 1$ and $2^i = n$.

In addition, using the fact that $r_\lambda = (2|\lambda\backslash\lambda_1| - 1)\cdot r_{\lambda\backslash\lambda_1}$ if λ has at least 2 parts, and the formula for z_λ, we easily obtain the recurrence:

$$S_n^{(i,j)} = \frac{(2(n-2^i)-1)^k}{2^i j} S_{n-2^i}^{(i,j-1)} \text{ for } (i,j)\in E_n \text{ with } 2^i < n,$$

valid for $j=1$ upon defining by convention $S_n^{(i,0)}$ as the sum of $S_n^{(i',j')}$ over all pairs $(i',j')\in E_n$ such that $i'<i$.

Thus in step (1), instead of directly drawing λ under $P(\lambda)$, we may first choose the pair (i,j) such that the largest part of λ is 2^i and has multiplicity j, that is, we draw $(i,j)\in E_n$ under distribution $P(i,j) = S_n^{(i,j)}/S_n$. Then we continue recursively at size $n' = n - 2^i j$, but conditioned on the largest part to be smaller than 2^i (that is, for the second step and similarly for later steps, we draw the pair (i',j') in $E_{n'}\cap\{i'<i\}$ under distribution $S_{n'}^{(i',j')}/S_{n'}^{(i,0)}$). Note that $|E_n| = \sum_{i\leq\log_2(n)}\lfloor n/2^i\rfloor = \Theta(n)$. Since we need all coefficients $S_m^{(i,j)}$ for $m\leq n$ and $(i,j)\in E_m$, we have to store $\Theta(n^2)$ coefficients. In addition it is easy to see (looking at the first expression in (4)) that each coefficient $S_m^{(i,j)}$ is a rational number of the form $a/m!$ with a an integer having $O(m\log(m))$ bits. Hence the overall storage bit-complexity is $O(n^3\log(n))$. About time complexity, starting at size n we first choose the pair (i,j) (with 2^i the largest part and j its multiplicity), which takes $O(|E_n|) = O(n)$ comparisons, and then we continue recursively at size $n - j\cdot 2^i$. At each step the choice of a pair (i,j) takes time $O(m)$ with $m\leq n$ the current size, and the number of steps is the number of distinct part-sizes in the finally output binary partition $\lambda\vdash n$. Since the number of distinct part-sizes in a binary partition of n is $O(\log(n))$, we conclude that the time complexity (in terms of the number of real-arithmetic comparisons) to draw λ is $O(n\log(n))$.

Acknowledgements. I thank Igor Pak for interesting discussions.

References

[1] F. Bergeron, G. Labelle, and P. Leroux. *Combinatorial Species and Tree-like Structures.* Cambridge University Press, 1997.

[2] S. Billey, M. Konvalinka, and F. Matsen IV. On the enumeration of tanglegrams and tangled chains. arXiv:1507.04976, 2015.

[3] I. Gessel. Counting tanglegrams with species. arXiv:1509.03867, 2015.

[4] K. Mahler. On a special functional equation. *Journal of the London Mathematical Society*, 1(2):115–123, 1940.

[5] F. Matsen IV, S. Billey, A. Kas, and M. Konvalinka. Tanglegrams: a reduction tool for mathematical phylogenetics. arXiv:1507.04784, 2015.

[6] George Pólya. Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. *Acta mathematica*, 68(1):145–254, 1937.

[7] J.-L. Rémy. Un procédé itératif de dénombrement d'arbres binaires et son application à leur génération aléatoire. *RAIRO, Informatique théorique*, 19(2):179–195, 1985.

[8] O. Rodrigues. Sur le nombre de manières de décomposer un polygone en triangles au moyen de diagonales. *Journal de Mathématiques Pures et Appliquées*, pages 547–548, 1838.