On the Combinatorics of Rooted Binary Phylogenetic Trees

Yun S. Song*

April 3, 2003

AMS Subject Classification: 05C05, 92D15

Abstract

We study *subtree-prune-and-regraft* (SPR) operations on leaf-labelled rooted binary trees, also known as rooted binary phylogenetic trees. This study is motivated by the problem of graphically representing evolutionary histories of biological sequences subject to recombination. We investigate some basic properties of the induced SPR-metric on the space $\mathscr{T}_n^{\mathrm{r}}$ of leaf-labelled rooted binary trees with n leaves. In contrast to the case of unrooted trees, the number $|U(T)|$ of trees in $\mathscr{T}_n^{\mathrm{r}}$ which are one SPR operation away from a given tree $T \in \mathscr{T}_n^{\mathrm{r}}$ depends on the topology of T. In this paper, we construct recursion relations which allow one to determine the unit-neighbourhood size $|U(T)|$ efficiently for any tree topology. In fact, using the recursion relations we are able to derive a simple closed-form formula for the unit-neighbourhood size. As a corollary, we construct sharp upper and lower bounds on the size of unit-neighbourhoods and investigate the diameter of $\mathscr{T}_n^{\mathrm{r}}$. Lastly, we consider an enumeration problem relevant to population genetics.

Keywords: rooted trees, ordered trees, subtree prune regraft, neighbourhood

1. Introduction

Biology abounds with examples where graphical representations and combinatorics have proved very useful. Through this bridge between biology and mathematics, many interesting ideas have been carried over from the latter and have lead to significant developments in the former. Of particular interest to geneticists is the usage of trees to represent evolutionary histories of biological sequences. In addition to obtaining a tree which best describes the evolutionary relationship of given sequences, one is often also interested in knowing how different a tree is from other trees; that is, one is interested in a quantitative measure of how far a tree is from another. The answer to that question, of course, depends on how the distance is measured, and therefore one needs to specify which *metric* should be used in measuring the distance between two trees.

A type of metric widely used in biology is that defined in terms of certain operations which rearrange trees [1, 5]; the distance between two trees is defined as the minimum

*Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK, song@stats.ox.ac.uk

number of operations required to transform one tree to the other. A particular kind of operation which will be the focus of this paper is the so-called *subtree pruning and regrafting* [5]. In a subtree-prune-and-regraft (SPR) operation, one detaches an edge from a tree T, thus "pruning" a subtree t from T, and "regrafts" t to somewhere else on the remaining part of T. We defer a more precise definition of SPR operations until §2.

In [1], Allen and Steel have considered the space $\mathscr{T}_n^{\mathrm{ur}}$ of leaf-labelled n-leaved *unrooted* binary trees. After having defined the neighbourhood of a tree $T \in \mathscr{T}_n^{\mathrm{ur}}$ as the set of all trees in $\mathscr{T}_n^{\mathrm{ur}}$ which are one SPR operation away from T, Allen and Steel have shown that the size of the neighbourhood of T does not depend on the topology of T and is equal to $2(n-3)(2n-7)$. Moreover, they have shown that the diameter $\mathrm{diam}_{SPR}(\mathscr{T}_n^{\mathrm{ur}})$, measured using the SPR-metric, satisfies the following bounds [1]:

$$n/2 - o(n) \leq \mathrm{diam}_{SPR}(\mathscr{T}_n^{\mathrm{ur}}) \leq n-3. \tag{1.1}$$

In the present paper, we investigate analogous questions for *rooted* trees, which, as we discuss presently, are more relevant to biology than *unrooted* trees. In contrast to the case of *unrooted* trees, the size $|U(T)|$ of the neighbourhood of a leaf-labelled *rooted* binary tree $T \in \mathscr{T}_n^{\mathrm{r}}$ depends on the topology of T. We are, however, able to construct recursion relations which can be used to compute $|U(T)|$ efficiently for any tree topology type. Furthermore, using the recursion relations, we derive a simple closed-form formula for $|U(T)|$. We find two particular topology types, one of which realises the maximum value of $|U(T)|$ and the other the minimum, and we combine this finding with the aforementioned results to construct sharp bounds for $|U(T)|$. Also, we show that the diameter of $\mathscr{T}_n^{\mathrm{r}}$ satisfies bounds similar to that shown in (1.1).

When representing genealogical processes by trees, it is natural to use rooted trees instead of unrooted trees, for the existence of a distinguished point on a tree enables us to define a sense of time direction; that is, time flows from the root to the leaves. This distinction between rooted and unrooted trees leads to observable differences in practice. For instance, in [2] Hein has proposed an algorithm for reconstructing the most parsimonious evolutionary histories of sequences which have undergone recombination. As he points out in the paper, if unrooted trees are used in the algorithm, internal contradictions might arise, thus preventing the construction of a graphical representation. If rooted trees are used, however, it could be possible to compute the exact minimum number of recombination events and thereby construct a consistent graphical representation [4]. Our findings from the present paper are used in [4], where SPR operations on rooted trees are used to represent recombination events.

In genetics, one is naturally lead to consider leaf-labelled rooted binary trees whose internal vertices, which correspond to biological events, are totally ordered. Such trees are sometimes called *ordered* trees. In this paper we consider an enumeration problem which arises in population genetics. Namely, we derive closed-form formulae for the number of rooted and ordered trees which are compatible with a bipartition of the label set.

This paper is organised as follows. In §2 we lay out some basic definitions and state a few fundamental results regarding the combinatorics of leaf-labelled rooted binary trees. In §3 we construct recursion relations for the size of unit-neighbourhoods and derive a closed-form formula for $|U(T)|$. In §4 we obtain sharp upper and lower bounds

on the size of unit-neighbourhoods. The diameter of the space $\mathscr{T}_n^{\mathrm{r}}$ is discussed in §5. In §6 we discuss the aforementioned enumeration problem relevant to genetics.

(NOTE: We have written a computer program to check explicitly all our results for $n \leq 9$.)

2. Preliminaries

2.1. Definitions

By a *rooted binary phylogenetic tree* we mean a leaf-labelled rooted binary tree whose branch lengths are not specified. The space of leaf-labelled rooted binary trees with n leaves is denoted by $\mathscr{T}_n^{\mathrm{r}}$. The degree of a vertex v is the number of edges which are incident with v. For $n \geq 2$, a tree in $\mathscr{T}_n^{\mathrm{r}}$ has n labelled degree-1 vertices called *leaves*; $n-2$ unlabelled degree-3 vertices; and a distinguished vertex of degree 2 called the *root*. A 1-leaved tree consists of a single labelled degree-0 vertex which serves as both the root and the leaf. A vertex which is not a leaf is called an *internal* vertex. The leaves of an n-leaved tree are bijectively labelled by a finite set $\mathscr{L}$ of n elements. Let $\mathscr{L}(T)$ be the label set for the leaves in $T \in \mathscr{T}_n^{\mathrm{r}}$. Then, for a subtree $s \subset T$, $\mathscr{L}(s) \subset \mathscr{L}(T)$ denotes the label set for the leaves in s. In the remainder of this paper, when we say a tree without any qualification, we shall mean a leaf-labelled rooted binary tree.

We say that two vertices $u, v \in T$ are *adjacent* if there exists an edge which joins u and v. A *path* from vertex v_0 to vertex v_k is an alternating sequence $v_0, e_1, v_1, e_2, \ldots, e_k, v_k$ of vertices v_i and edges e_i, such that (1) e_i joins v_{i-1} and v_i, and (2) all e_is and v_is are distinct. For v a degree-3 vertex in T, we define $\gamma(v)$ as the number of degree-3 vertices, not including v itself, in the path between v and the root of T.

In a rooted tree, time flows from the root to the leaves. We say that vertex $v \in T$ is a *descendant* of vertex $u \in T$ if there exists a path from u to v which goes strictly forward in time; u is called an *ancestor* of v. A *subtree* s of a tree $T \in \mathscr{T}_n^{\mathrm{r}}$ is a tree in $\mathscr{T}_{n'}^{\mathrm{r}}$, where $n' \leq n$, and is defined by the property that if a vertex $v \in T$ is contained in s, then so are all its descendants. In this paper a subtree whose root is adjacent to the root of T is called an R-subtree.

An n-leaved rooted binary tree contains $2n-2$ edges. For any (sub)tree s, we denote by $\ell(s)$ the number of leaves in s and define $\eta(s) := 2\ell(s) - 2$, which is equal to the number of edges in s.

Lemma 2.1. (Schröder) *The number of inequivalent leaf-labelled rooted binary trees with n leaves is [3]*

$$R(n) := |\mathscr{T}_n^{\mathrm{r}}| = (2n-3)!! = (2n-3) \times (2n-5) \times \cdots \times 3 \times 1 = \frac{(2n-2)!}{2^{n-1}(n-1)!}.$$

2.2. SPR Operations

There are three kinds of SPR operations that can be performed on leaf-labelled rooted binary trees. An illustration of these operations is shown in Figure 1. In what follows, let T (resp. T') denote a tree before (resp. after) an SPR operation. The notation $T \setminus t$

denotes the part of T obtained from removing a subtree t and the edge incident with the root of t but not contained in t. In words the three SPR operations are as follows.

(1) An edge e is cut to prune a non-R-subtree t, and t is regrafted onto a pre-existing edge in the remaining part $T \setminus t$ of T, thus creating a new degree-3 vertex. The vertex in $T \setminus t$ where e used to be incident gets removed. The root of T remains the root of T'. (In Figure 1, $T \to T_1$ is an example of this kind. The edge e_b is cut and then regrafted onto the edge e_a.)

(2) Let s_1 and s_2 be the two R-subtrees of T, and let e_1 and e_2, respectively, be the edges which join their roots to the root of T. The edge e_1 is cut to prune s_1, and s_1 is regrafted onto a pre-existing edge in s_2. The edge e_2 gets removed and the degree-3 vertex in s_2 where e_2 used to be incident gets replaced by a degree-2 vertex, which becomes the root of T'. (In Figure 1, $T \to T_2$ is an example of this kind. The edge e_c is cut and regrafted onto e_a. The root of the R-subtree containing t_1, t_2 and t_3 becomes the root of T_2.)

(3) An edge e is cut to prune a non-R-subtree t, and t is joined to the root of T. The root of T' is given by creating a new vertex of degree 2 on e. (In Figure 1, $T \to T_3$ is an example of this kind. The edge e_b is cut and then joined to the root of T. A new degree-2 vertex is created on the edge and it serves as the root of T_3.)

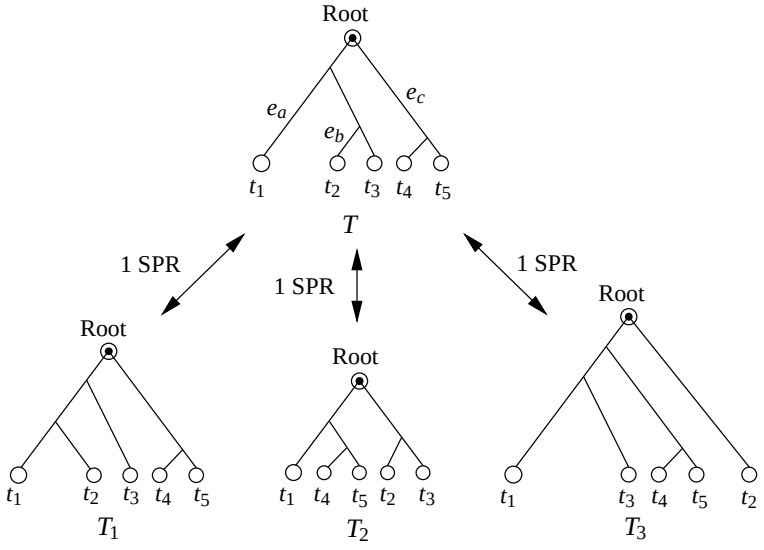

Figure 1: An illustration of SPR operations. Big open circles ◯ labelled by t_j represent subtrees.

For any pair of trees $T, T' \in \mathscr{T}_n^{\mathrm{r}}$, we measure the distance between them using the SPR-metric $d : \mathscr{T}_n^{\mathrm{r}} \times \mathscr{T}_n^{\mathrm{r}} \to \mathbb{Z}_{\geq 0}$; that is, the distance $d(T, T')$ is a non-negative integer defined as the minimum number of SPR operations necessary to transform T into T'.

3. The Unit-Neighbourhood of a Tree

We define the unit-neighbourhood of a tree $T \in \mathscr{T}_n^{\mathrm{r}}$ as

$$U(T) = \{T' \in \mathscr{T}_n^{\mathrm{r}} \,|\, d(T,T') = 1\}.$$

3.1. Topology Types

We here define two topology types, shown in Figure 2, for which simple recursion relations for $|U(T)|$ will later be formulated. A type A tree is characterised by the feature that only a single leaf is on one side of the root. In Figure 2(b), if v denotes the degree-3 vertex with which the leaf l_n is adjacent, then $k = \gamma(v)$. The notion of "left" and "right" in the figure is irrelevant. For ease of reference, we have given labels to some edges; these labels are not a part of the definition of a leaf-labelled rooted binary tree. The reader should refer to the captions therein for further explanation. We emphasise that what we introduce here does *not* define a classification, since a tree can fall into more than one type. For instance, for $n \geq 3$, any tree $T \in \mathscr{T}_n^{\mathrm{r}}$ is type B, but it may also be type A.

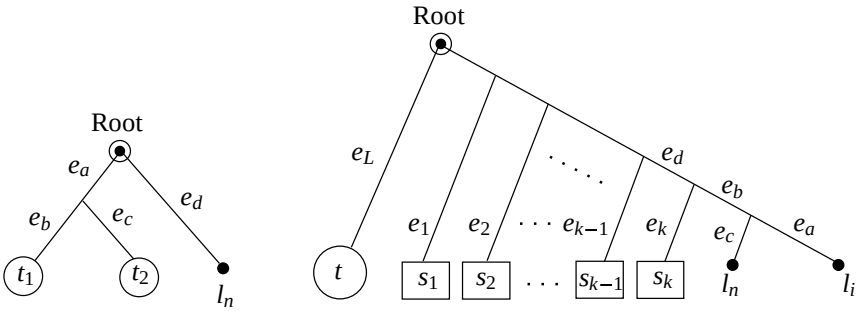

Figure 2: A schematic representation of topology types. Big open circles ◯ and boxes □ represent subtrees, and every subtree contains at least one leaf. Leaves are labelled by $l_n, l_i \in \mathscr{L}$. (a) An n-leaved type A tree. (b) An n-leaved type B tree.

3.2. "$-$" Operations on Trees

We here define a reduction operation which will be used in our recursion relations. The "$-$" operation we presently define reduces the number of leaves in a tree by one. In the following discussion, we use the labels shown in Figure 2.

(1) If $T \in \mathscr{T}_n^{\mathrm{r}}$ is type A, then $T - l_n$ is given by removing the leaf l_n; removing the root of T; removing the edges e_a and e_d; and making the vertex where e_a, e_b, e_c used to be incident into the root of $T - l_n$. An example of this kind of operation is illustrated in Figure 3(a).

(2) If $T \in \mathscr{T}_n^{\mathrm{r}}$ is type B, then $T - l_n$ is given by removing the leaf l_n; removing the edge e_c; removing the degree-3 vertex where e_a, e_b, e_c used to be incident; and merging the edges e_a and e_b into one. An example of this kind of operation is illustrated in Figure 3(b).

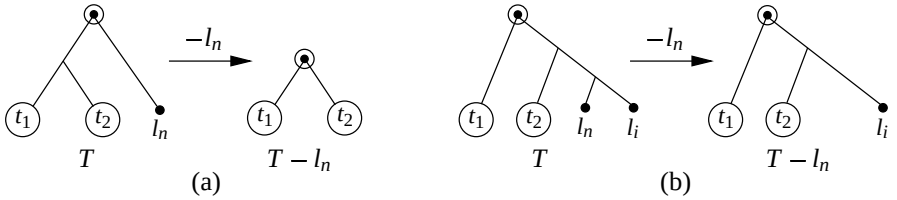

Figure 3: An illustration of "$-$" operations. (a) T is type A. (b) T is type B, with $k = 1$.

3.3. Recursion Relations for $|U(T)|$

In this subsection, we construct recursion relations for $|U(T)|$. The topology types defined in §3.1 constitute a rather coarse description. For instance, type B encompasses many distinct tree topologies. It is interesting to note that the dependence of type B recursion relation on tree topology is encoded in a single parameter k.

The recursion relations can be applied in several different ways, depending on which type one chooses to call a tree and which "$-$" operation one chooses to use. The final answer for $|U(T)|$, however, does not depend on how one chooses to compute it. In fact, type B recursion relation alone is sufficient for computing $|U(T)|$ for any tree T. Type A recursion relation, however, will be useful for our discussion in §4.

Proposition 3.2. *Let $n \geq 4$ and let $T \in \mathscr{T}_n^{\mathrm{r}}$. Then, the size of the unit-neighbourhood $U(T)$ satisfies the recursion relation*

$$|U(T)| = \begin{cases} |U(T - l_n)| + 6n - 16\,, & \text{if } T \text{ is type A,} \\ |U(T - l_n)| + 2(4n - k - 11)\,, & \text{if } T \text{ is type B,} \end{cases}$$

where l_n is as shown in Figure 2, and $T - l_n$ is an $(n-1)$-leaved tree obtained using the "$-$" operation defined in §3.2. In the recursion relation for type B trees, k is a non-negative integer defined as in Figure 2(b).

REMARK: $|U(T)| = 2$, for all $T \in \mathscr{T}_3^{\mathrm{r}}$, serves as the boundary condition for the recursion relations.

Proof. We have divided our proof into several parts. In our discussion, we shall conform to the notations shown in Figure 2.

TYPE A:

(A-1) A tree in $U(T)$ can be generated by a single SPR operation within the part to the left of the root of T. There are $|U(T - l_n)|$ such operations.

(A-2) Any edge except for e_a and e_d can be detached from T and regrafted onto e_d to generate a tree in $U(T)$ which has not been included in (A-1). There are $\eta(T)-2 = 2n-4$ such edges in T.

(A-3) The edge e_d can be detached and regrafted onto an edge other than e_a, e_b, e_c and e_d to generate a tree in $U(T)$ which has not been included in (A-1) or (A-2). There are $\eta(T)-4 = 2n-6$ possibilities.

(A-4) Any edge except for e_a, e_b, e_c and e_d can be detached from T and regrafted onto the root of T to generate a tree in $U(T)$ which has not been included above. There are $\eta(T)-4 = 2n-6$ such edges.

Adding up the contributions gives

$$|U(T)| = |U(T-l_n)| + (2n-4) + (2n-6) + (2n-6) = |U(T-l_n)| + 6n - 16\,.$$

TYPE B ($k \geq 1$):

(B-1) There are $|U(T-l_n)|$ inequivalent SPR operations which do not directly involve e_a or e_c (i.e. neither cutting them nor regrafting onto them).

(B-2) The edge e_a can be detached from T and regrafted onto any edge except for e_a, e_b and e_c to yield a new tree in $U(T)$. There are $\eta(T)-3 = 2n-5$ inequivalent such SPR operations.

(B-3) The edge e_c can be detached from T and regrafted onto any edge except for e_a, e_b, e_c, e_d, e_k to generate a tree in $U(T)$ which has not already been accounted for in (B-1) or (B-2). There are $2n-2-5 = 2n-7$ such SPR operations.

(B-4) Any edge in the subtree t can be detached from t and regrafted onto either e_a or e_c. There are $2\eta(t)$ such operations which generate inequivalent trees in $U(T)$.

(B-5) For $j = 1, 2, \ldots, k-1$, if $\ell(s_j) > 1$, any edge in the subtree s_j can be detached from s_j and regrafted onto either e_a or e_c. Also, e_j can be detached from T and regrafted onto either e_a or e_c. There are a total of $2\sum_{j=1}^{k-1}[\eta(s_j)+1]$ inequivalent such SPR operations.

(B-6) If $\ell(s_k) > 1$, any edge in the subtree s_k can be detached from s_k and regrafted onto either e_a or e_c. There are $2\eta(s_k)$ such SPR operations. (Note: detaching e_k and regrafting it onto either e_a or e_c generates a tree already included in (B-2).)

(B-7) The edge e_a or e_c can be detached and regrafted onto the root. This contributes 2 to $|U(T)|$.

(B-8) The edge e_L can be detached and regrafted onto either e_a or e_c to yield a tree in $U(T)$ which has not already been included above. This contributes 2 to $|U(T)|$.

In summary, we have

$$\begin{aligned}|U(T)| &= |U(T-l_n)| + (2n-5) + (2n-7) + 2\eta(t) + 2\sum_{j=1}^{k-1}[\eta(s_j)+1] + 2\eta(s_k) + 4 \\ &= |U(T-l_n)| + 2\left[2(n-3) + \eta(t) + \sum_{j=1}^{k}\eta(s_j) + (k-1) + 2\right] \\ &= |U(T-l_n)| + 2[4n-k-11]\,,\end{aligned}$$

where the last line follows from $\eta(t) + \sum_{j=1}^{k}\eta(s_j) = 2n-2k-6$.

TYPE B ($k=0$):
In this case, the only edges to the right of the root are e_a, e_b and e_c (c.f. Figure 2(b)).

(B-1′) There are $|U(T-l_n)|$ inequivalent SPR operations which do not involve e_a or e_c directly.

(B-2′) An edge in the subtree t can be detached from t and regrafted onto either e_a or e_c. There are $2\eta(t)$ inequivalent such operations.

(B-3′) The edge e_c can be detached from T and regrafted onto any edge to the left of the root. There are $\eta(T)-3 = 2n-5$ such operations.

(B-4′) Likewise, the edge e_a can be detached from T and regrafted onto any edge to the left of the root. Again, there are $\eta(T)-3 = 2n-5$ such operations.

Note that, since $k=0$, detaching e_a (resp. e_c) and regrafting it to the root of T is equivalent to detaching e_c (resp. e_a) and regrafting it to e_L. Also, detaching e_L and regrafting it to e_a (resp. e_c) is equivalent to detaching e_a (resp. e_c) and regrafting it to e_L. These operations generate trees which have already been included in the above list. Hence, we obtain

$$|U(T)| = |U(T-l_n)| + 2\eta(t) + 2(2n-5) = |U(T-l_n)| + 2(4n-11)\,,$$

where $\eta(t) = 2(n-2)-2 = 2n-6$ has been used. ∎

3.4. A Closed-Form Formula for $|U(T)|$

As we have mentioned before, it is always possible to compute $|U(T)|$ only using the type B recursion relation. Applying the recursion relation in a systematic way, one can obtain the following result:

Proposition 3.3. *Let $n \geq 3$ and let $T \in \mathscr{T}_n^{\mathrm{r}}$. Let $\{v_1, v_2, \ldots, v_{n-2}\}$ be the set of degree-3 vertices in T. Then, with $\gamma(v_i)$ defined as in §2.1, $|U(T)|$ is given by*

$$|U(T)| = 2(n-2)(2n-5) - 2\sum_{i=1}^{n-2}\gamma(v_i)\,. \tag{3.2}$$

Proof. Label the degree-3 vertices of T by $\{v_1, v_2, \ldots, v_{n-2}\}$ so that

$$\gamma(v_{n-2}) \leq \gamma(v_{n-3}) \leq \cdots \leq \gamma(v_1)\,. \tag{3.3}$$

Now, successively perform "$-$" operations in the order

$$-l_1\,,\ -l_2\,,\ \ldots\,,\ -l_{n-3}\,, \tag{3.4}$$

where l_i denotes a leaf incident with v_i in $T - \sum_{k=1}^{i-1} l_k$. Note that, because of the imposed ordering in (3.3), "$-l_i$" operation does not change the value of $\gamma(v_j)$ for $j = i+1, i+2, \ldots, n-2$. After performing all the operations in (3.4), we end up with a 3-leaved tree, whose unit-neighbourhood size is 2. In summary, using Proposition 3.2 we obtain

$$|U(T)| \;=\; 2 + \sum_{m=4}^{n} 2(4m-11) - 2\sum_{i=1}^{n-3} \gamma(v_i) \;=\; 2(n-2)(2n-5) - 2\sum_{i=1}^{n-2} \gamma(v_i)\,,$$

where in the last equality $\gamma(v_{n-2})$ has been added to the summation (Note that the ordering in (3.3) implies that $\gamma(v_{n-2}) = 0$, so adding it to the summation does not change the value of $|U(T)|$). ∎

Recall Allen and Steel's formula [1]

$$AS(n) := 2(n-3)(2n-7)$$

for the size of the unit-neighbourhood of an *unrooted* binary tree in $\mathscr{T}_n^{\mathrm{ur}}$. The first term in our formula (3.2) is none other than $AS(n+1)$. This result reflects the fact that there exists a one-to-one correspondence between the set $\mathscr{T}_{n+1}^{\mathrm{ur}}$ of leaf-labelled *unrooted* binary trees with $n+1$ leaves and the set $\mathscr{T}_n^{\mathrm{r}}$ of leaf-labelled *rooted* binary trees with n leaves. Furthermore, the definition of an SPR operation for n-leaved *rooted* trees is more restrictive than that for $(n+1)$-leaved *unrooted* trees. That is, there are more inequivalent SPR operations for $(n+1)$-leaved unrooted trees than for n-leaved rooted trees. The second term in (3.2) is the necessary correction term which accounts for this fact.

4. Sharp Bounds on the Size of Unit-Neighbourhoods

In this section, we define two special types of trees and examine the size of their unit-neighbourhoods. We then use our findings to obtain sharp upper and lower bounds for $|U(T)|$.

4.1. Two Special Types of Trees

Consider the sequence $a_1, a_2, a_3, \ldots$, where

$$a_m = \lfloor \log_2(m+1) - 1 \rfloor\,. \tag{4.5}$$

Here, $\lfloor \cdot \rfloor$ denotes the greatest integer function, also known as the floor function. More explicitly, the sequence is of the form

$$0,0,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,4,\ldots\,,$$

containing two 0s, four 1s, eight 2s, sixteen 3s, thirty two 4s, etc.

Also, consider the strictly ascending sequence $b_1, b_2, b_3, \ldots$, where $b_m = m-1$; that is, the sequence is

$$0, 1, 2, 3, 4, 5, 6, 7, \ldots,$$

containing one 0, one 1, one 2, one 3, one 4, etc.

Let $T \in \mathscr{T}_n^{\mathrm{r}}$. If its $n-2$ degree-3 vertices can be labelled by $\{v_1, v_2, \ldots, v_{n-2}\}$ so that $\gamma(v_m) = a_m$ (resp. $\gamma(v_m) = b_m$) for every $m \in \{1, 2, \ldots, n-2\}$, then we shall call T a "γ-exponential" (resp. "γ-uniform") tree. Examples of γ-exponential and γ-uniform trees are shown in Figure 4.

(SIDE REMARK: The name "γ-exponential" is derived from the fact that non-negative integers are *exponentially* distributed in $\{a_1, a_2, \ldots, \}$, and the name "γ-uniform" from the fact that non-negative integers are *uniformly* distributed in $\{b_1, b_2, \ldots, \}$.)

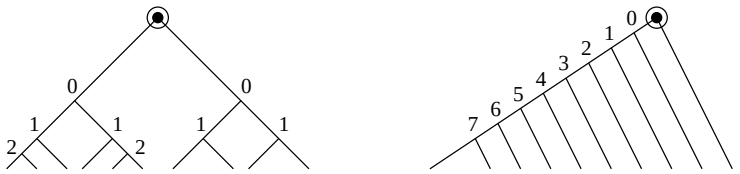

Figure 4: Examples of special types of trees. Leaf labels are suppressed and the values of $\gamma(v_i)$ for degree-3 vertices v_i are shown. (a) A 10-leaved γ-exponential tree. (b) A 10-leaved γ-uniform tree.

4.2. Unit-Neighbourhoods of γ-Exponential and γ-Uniform Trees

In this subsection, we examine the size of unit-neighbourhoods of γ-exponential and γ-uniform trees. As we show in the following proposition, these trees are special in the sense that they realise extreme values of the unit-neighbourhood size.

Proposition 4.4. *Let $n \geq 4$ and define $\delta_{\max}(n) := \max_{T \in \mathscr{T}_n^{\mathrm{r}}} |U(T)|$ and $\delta_{\min}(n) := \min_{T \in \mathscr{T}_n^{\mathrm{r}}} |U(T)|$. Then,*

I. *$|U(T)| = \delta_{\max}(n)$ if and only if T is γ-exponential,*

II. *$|U(T)| = \delta_{\min}(n)$ if and only if T is γ-uniform.*

Proof. Consider the case $n = 4$. It turns out that every tree $T \in \mathscr{T}_4^{\mathrm{r}}$ is either γ-exponential or γ-uniform. Moreover, we can use Proposition 3.2 or Proposition 3.3 to show explicitly that $|U(T)| = 12$ if T is γ-exponential, whereas $|U(T)| = 10$ if T is γ-uniform. Hence, both statements I and II in the proposition are true for $n = 4$.

PROOF OF PART I: Let $\mathscr{H}_{\mathrm{I}}$ denote the induction hypothesis that, for $4 \leq n \leq r-1$ where $r \geq 5$, $|U(T)| = \delta_{\max}(n)$ if T is γ-exponential. Let T be an r-leaved γ-exponential tree and let l_r denote a leaf adjacent to an internal vertex v with $\gamma(v) = a_{r-2}$, where a_{r-2} is defined in (4.5). Then, $T - l_r$ also is a γ-exponential tree. Hence, by our induction

hypothesis $\mathcal{H}_{\rm I}$, we know that $|U(T-l_r)| = \delta_{\max}(r-1)$. Moreover, since a γ-exponential tree is type B for $r \geq 4$, we can use the recursion relation from Proposition 3.2 to obtain

$$|U(T)| = \delta_{\max}(r-1) + 2(4r - a_{r-2} - 11)\,.$$

Suppose $T' \in \mathscr{T}_r^{\rm r}$ is type A. Then, $|U(T'-l')| \leq \delta_{\max}(r-1)$, where l' is the single leaf on the right hand side of the root in Figure 2(a). Furthermore, since $6r-16 < 2(4r-a_{r-2}-11)$, for all $r \geq 5$, we conclude that $|U(T')| = |U(T'-l')| + 6r - 16 < |U(T)|$.

We now show that for all type B tree $T' \in \mathscr{T}_r^{\rm r}$ which is not γ-exponential, $|U(T')| < |U(T)|$. Let l' be a leaf in T' such that k' is as large as it can be in the following formula from Proposition 3.2: $|U(T')| = |U(T'-l')| + 2(4r-k'-11)$. By definition, $\delta_{\max}(r-1) \geq |U(T'-l')|$. Furthermore, since T' is not γ-exponential, $k' > a_{r-2}$ for all $r \geq 5$, and therefore $2(4r-a_{r-2}-11) > 2(4r-k'-11)$. Hence, $|U(T')| < |U(T)|$, and we thus conclude that if T is γ-exponential, then $|U(T)| = \delta_{\max}(r)$. This completes our induction.

The converse can be shown as follows. Let T be an n-leaved tree such that $|U(T)| = \delta_{\max}(n)$. Then, from the formula for $|U(T)|$ in (3.2), we know that $\sum_{i=1}^{n-2}\gamma(v_i)$ must be as small as possible. But, in a rooted binary tree, $\gamma(v_1),\gamma(v_2),\ldots,\gamma(v_{n-2}) = a_1,a_2,\ldots,a_{n-2}$ gives the minimum value of $\sum_{i=1}^{n-2}\gamma(v_i)$, and therefore T must be γ-exponential.

PROOF OF PART II: Assume that, for $4 \leq n \leq r-1$ where $r \geq 5$, $|U(T)| = \delta_{\min}(n)$ if T is γ-uniform. We refer to this assumption as hypothesis $\mathcal{H}_{\rm II}$. Let T be an r-leaved γ-uniform tree.

Since a γ-uniform tree is of type A, we can use the type A recursion relation from Proposition 3.2 to obtain

$$|U(T)| = |U(T-l_r)| + 6r - 16 = \delta_{\min}(r-1) + 6r - 16,$$

where the second equality follows from the induction hypothesis $\mathcal{H}_{\rm II}$.

Suppose T' is type B. Then, $|U(T')| = |U(T'-l'_r)| + 2(4r-k-11) \geq \delta_{\min}(r-1) + 2(4r-k-11)$, where k is as shown in Figure 2(b). But, in a tree with r leaves, k is bounded from above. More precisely, $k \leq r-3$, and we therefore have $6r-16 \leq 2(4r-k-11)$ for every $r \geq 5$. Hence, we conclude that $|U(T')| \geq |U(T)|$.

Suppose T' is a type A tree which is not γ-uniform and let l'_r denote the single leaf on the right hand side of the root in Figure 2(a). Then, applying the type A recursion relation gives $|U(T')| = |U(T'-l'_r)| + 6r - 16 \geq \delta_{\min}(r-1) + 6r - 16$. Hence, $|U(T')| \geq |U(T)|$. We have thus shown that if T is an r-leaved γ-uniform tree, then $|U(T)| = \delta_{\min}(r)$. This completes our induction.

We now sketch the proof of the converse. Let T be an n-leaved tree such that $|U(T)| = \delta_{\min}(n)$. Then, it implies that $\sum_{i=1}^{n-2}\gamma(v_i)$ must be as large as possible in (3.2). It is easy to show that, in a rooted binary tree, $\gamma(v_1),\gamma(v_2),\ldots,\gamma(v_{n-2}) = 0,1,2,\ldots,n-3$ yields the maximum value of $\sum_{i=1}^{n-2}\gamma(v_i)$. Thus, T must be γ-uniform. ∎

4.3. The Bounds

Using the recursion relations from §3.3 and the results from §4.2, we can derive a corollary of the following form:

Corollary 4.5. *For all $T \in \mathscr{T}_n^{\mathrm{r}}$, $n \geq 4$, the size of the unit-neighbourhood $U(T)$ satisfies the bounds*

$$3n^2 - 13n + 14 \leq |U(T)| \leq 4(n-2)^2 - 2\sum_{m=1}^{n-2} \lfloor \log_2(m+1) \rfloor .$$

That is, $\delta_{\min}(n) = 3n^2 - 13n + 14$ *and* $\delta_{\max}(n) = 4(n-2)^2 - 2\sum_{m=1}^{n-2} \lfloor \log_2(m+1) \rfloor$.

Proof. Let T be an n-leaved γ-uniform tree. Then, by Proposition 3.3 and Proposition 4.4, we have

$$|U(T)| = \delta_{\min}(n) = 2(n-2)(2n-5) - 2\sum_{m=1}^{n-2} b_m,$$

where $b_m = m-1$ (c.f. §4.1). Hence, $\delta_{\min}(n) = 2(n-2)(2n-5) - 2(n-3)(n-2)/2 = 3n^2 - 13n + 14$.

Let T be an n-leaved γ-exponential tree. Then, it follows from Proposition 3.3 and Proposition 4.4 that

$$|U(T)| = \delta_{\max}(n) = 2(n-2)(2n-5) - 2\sum_{m=1}^{n-2} a_m \;=\; 4(n-2)^2 - 2\sum_{m=1}^{n-2} \lfloor \log_2(m+1) \rfloor ,$$

and we have our desired result. ∎

5. Diameter of $\mathscr{T}_n^{\mathrm{r}}$

As Allen and Steel have done for unrooted trees [1], we can obtain the following result for rooted trees:

Proposition 5.6. *Let $n \geq 3$ and let* $\mathrm{diam}_{SPR}(\mathscr{T}_n^{\mathrm{r}})$ *denote the diameter of $\mathscr{T}_n^{\mathrm{r}}$, defined as the maximum value of $d(T,T')$ over all trees $T, T' \in \mathscr{T}_n^{\mathrm{r}}$. Then,*

$$n/2 - o(n) \leq \mathrm{diam}_{SPR}(\mathscr{T}_n^{\mathrm{r}}) \leq n-2.$$

Proof. From Corollary 4.5, we know that $\delta_{\max}(n) < 4(n-2)^2$. Following exactly the same line of reasoning as Allan and Steel have done in [1], one can analyse

$$[4(n-2)^2]^{\mathrm{diam}_{SPR}(\mathscr{T}_n^{\mathrm{r}})} \geq (2n-3)!!$$

using Stirling's approximation to derive the lower bound. The lower bound for the rooted case is the same as that in the unrooted case, because for both cases the unit-neighbourhood size grows quadratically with respect to n.

For small values of n, say $n \leq 6$, it is easy to come up with examples of $T, T' \in \mathscr{T}_n^{\mathrm{r}}$ such that $d(T,T') = n-2$. We wish to show that, for all $n \geq 3$, $n-2$ SPR operations are sufficient to transform any $T_1 \in \mathscr{T}_n^{\mathrm{r}}$ to any $T_2 \in \mathscr{T}_n^{\mathrm{r}}$. Note that the root of a tree partitions the label set $\mathscr{L}$ into two disjoint proper subsets of $\mathscr{L}$. Let $\{A_1, A_1^c\}$ and $\{A_2, A_2^c\}$ be such bipartitions of $\mathscr{L}$ associated to two trees T_1 and T_2, respectively. Here, A_i denotes a

proper subset of $\mathcal{L}$ and A_i^c its complement relative to $\mathcal{L}$. Let $S_1 \in \{A_1, A_1^c\}$ and $S_2 \in \{A_2, A_2^c\}$. Since A_1, A_1^c, A_2 and A_2^c are proper subsets, if $S_1 \cap S_2 = \varnothing$, then $S_1^c \cap S_2 \neq \varnothing$ and $S_1 \cap S_2^c \neq \varnothing$. Therefore, it is always possible to label the bipartitions so that $A_1 \cap A_2 \neq \varnothing$ and $A_1^c \cap A_2^c \neq \varnothing$. Upon making such a choice of labelling, let $l_i \in A_1 \cap A_2$ and $l_j \in A_1^c \cap A_2^c$. Note that the leaves l_i and l_j are on opposite sides of the root in both T_1 and T_2. Now, prune all the leaves, except for l_i and l_j, from T_1 and then regraft those $n-2$ leaves, labelled by $\mathcal{L} \setminus \{l_i, l_j\}$, to make T_2. It is clear that this is always possible. Thus we conclude that $d(T_1, T_2) \leq (n-2)$ for all $T_1, T_2 \in \mathscr{T}_n^{\mathrm{r}}$. ∎

6. Number of Trees Compatible with a Bipartition of $\mathcal{L}$

The set $\{v_1, v_2, \ldots, v_{n-2}\}$ of degree-3 vertices in $T \in \mathscr{T}_n^{\mathrm{r}}$ is a partially ordered set whose binary relation denoted $\leq$ is given by ancestral relation; we say that $v_i < v_j$ if v_i is a descendant of v_j. Two degree-3 vertices v_i and v_j are *incomparable* if v_i is not in the path to the root from v_j and vice versa. An *ordered* tree is a leaf-labelled rooted binary tree whose corresponding set $\{v_1, v_2, \ldots, v_{n-2}\}$ of degree-3 vertices is a *totally* ordered set; that is, for any two vertices v_i and v_j, either $v_i < v_j$ or $v_j < v_i$. In this case, the binary relation is given by age ordering. As before, $v_i < v_j$ if v_i is a descendant of v_j. If there exists no ancestral relation between v_i and v_j, then either $v_i < v_j$ or $v_j < v_i$ is allowed. Furthermore, we impose the condition that $v_i \neq v_j$ if $i \neq j$. Two trees equivalent as leaf-labelled rooted binary trees are distinct as ordered trees if the ordering of their degree-3 vertices are different. It is well-known in population genetics that the number of inequivalent ordered trees with n leaves is

$$D(n) := \prod_{m=2}^{n} \binom{m}{2} = \frac{n!(n-1)!}{2^{n-1}}.$$

Recall that $R(n) := |\mathscr{T}_n^{\mathrm{r}}| = (2n-3)!!$.

Let $\{B, B^c\}$ denote a bipartition of the label set $\mathcal{L}$ into two proper subsets. A tree T is said to be *compatible* with the bipartition $\{B, B^c\}$ if there exists an edge in T such that cutting the edge decomposes T into two connected components, one containing the leaves labelled by B and the other the leaves labelled by B^c. In population genetics – for example, when using the so-called infinite-sites model – the number of trees compatible with a bipartition of $\mathcal{L}$ is a quantity of interest. Suppose $|B| = k$ and $|B^c| = n-k$, and let $w^{\mathrm{r}}(n,k)$ (resp. $w^{\mathrm{o}}(n,k)$) denote the number of rooted trees (resp. ordered trees) compatible with the bipartition $\{B, B^c\}$. Clearly, if $k = 1$ or $k = n-1$, then $w^{\mathrm{r}}(n,k) = R(n)$ and $w^{\mathrm{o}}(n,k) = D(n)$. For $2 \leq k \leq n-2$, it is not difficult to show that the number of rooted trees compatible with $\{B, B^c\}$ is

$$w^{\mathrm{r}}(n,k) := (2n-3)\, R(k)\, R(n-k).$$

For ordered trees, we have the following result:

Proposition 6.7. *For $n \geq 4$ and $2 \leq k \leq n-2$, the number of ordered trees compatible with the bipartition $\{B, B^c\}$, where $|B| = k$ and $|B^c| = n-k$, is*

$$w^{\mathrm{o}}(n,k) := D(k)D(n-k)\left[\binom{n}{k-1} + \binom{n}{n-k-1} - \binom{n-2}{k-1}\right].$$

Proof. We first show that $w^{\circ}(n,k)$ is given by the following expression:

$$\begin{aligned} w^{\circ}(n,k) \;:=\; & D(k)\sum_{p=0}^{n-k-1}\left[D(n-k-p+1)\binom{p+k-2}{p}\prod_{s=0}^{p-1}\binom{n-k-s}{2}\right] \\ & + D(n-k)\sum_{p=0}^{k-1}\left[D(k-p+1)\binom{p+n-k-2}{p}\prod_{s=0}^{p-1}\binom{k-s}{2}\right] \\ & - D(k)D(n-k)\binom{n-2}{k-1}\,. \end{aligned} \tag{6.6}$$

Consider an urn containing k black balls labelled by B and $n-k$ white balls labelled by B^c. Draw two balls from the urn. If one black ball and one white ball are drawn, then simply replace both balls back into the urn. If two black (resp. white) balls labelled X_i and X_j are drawn, then replace with a single black (resp. white) ball labelled $X_i \oplus X_j$. Here, X_i could be $l_{i_1} \oplus l_{i_2} \oplus \cdots \oplus l_{i_j}$, where $l_{i_1}, l_{i_2}, \ldots, l_{i_j} \in \mathcal{L}$. Note that, in total, $k-1$ pairs of black balls can be drawn. When the $(k-1)^{\text{th}}$ pair of *black* balls labelled X_i and X_j are drawn, then replace with a *white* ball labelled $X_i \oplus X_j$. If only white balls remain in the urn, keep drawing pairs and replace with a white ball with a new label until only one white ball remains in the urn. Every possible sequence of draws ends up with a single white ball labelled $l_1 \oplus l_2 \oplus \cdots \oplus l_n$, where $\{l_1, l_2, \ldots, l_n\} = \mathcal{L}$.

There exists a one-to-one correspondence between the set of sequences of distinct urn contents which arise in the above urn model and the set of n-leaved ordered trees which contain an ordered subtree with k leaves labelled by B. The ordering of urn contents in a sequence corresponds to the ordering of internal vertices in an ordered tree. The initial set of balls correspond to the leaves and a ball with a composite label $l_{i_1} \oplus l_{i_2} \oplus \cdots \oplus l_{i_j}$ corresponds to an internal vertex whose descendant leaves are precisely $l_{i_1}, l_{i_2}, \cdots, l_{i_j}$. The white ball replaced into the urn when the $(k-1)^{\text{th}}$ pair of black balls are drawn corresponds to the root of the k-leaved ordered subtree whose leaves are labelled by B.

When the $(k-1)^{\text{th}}$ pair of black balls are drawn, we can associate to the completed sequence of black ball contents an ordered tree with k leaves labelled by B. By analogy, there are $D(k)$ inequivalent ways of achieving this. Suppose p pairs of white balls have been drawn before the $(k-1)^{\text{th}}$ pair of black balls are drawn. Since all balls are labelled, there are $\prod_{s=0}^{p-1}\binom{n-k-s}{2}$ inequivalent ways of drawing p pairs of white balls. Moreover, a sequence $\alpha_1, \alpha_2, \ldots, \alpha_p$ of p pairs of white balls drawn and a sequence $\beta_1, \beta_2, \ldots, \beta_{k-2}$ of $k-2$ pairs of black balls drawn can be combined in $\binom{p+k-2}{p}$ inequivalent ways to form a longer sequence of length $p+k-2$ while maintaining the ordering of α_is and that of β_is. Recall that, when the $(k-1)^{\text{th}}$ pair of black balls are drawn, a labelled white ball is replaced into the urn. Hence, the number of white balls then remaining in the urn is $n-k-p+1$. Subsequent drawings of pairs of white balls with replacement correspond to generating $D(n-k-p+1)$ ordered trees. Lastly, we note that p can take any value between 0 and $n-k-1$, inclusive. We have thus derived the expression in the first line of (6.6).

The expression in the second line of (6.6) is obtained by replacing k with $n-k$, and vice versa, in the above paragraphs. It corresponds to the number of n-leaved ordered trees which contain an ordered subtree with $n-k$ leaves labelled by B^c.

The last line in (6.6) corrects for double counting. That is, both the first line and the second line in (6.6) include the number of ordered trees which contain both an ordered subtree with k leaves labelled by B and an ordered subtree with $n-k$ leaves labelled by B^c. The combinatorial factor $\binom{n-2}{k-1}$ is for the ordering of the $n-k-1$ internal vertices of an $(n-k)$-leaved ordered tree relative to the $k-1$ internal vertices of a k-leaved ordered tree. This completes the derivation of (6.6).

Now, notice that $\prod_{s=0}^{j}\binom{i-s}{2} = D(i)/D(i-j-1)$, which implies that the expression inside the bracket in the first line of (6.6) is equal to $D(n-k)\binom{n-k-p+1}{2}\binom{p+k-2}{p}$. Furthermore, since $\sum_{p=0}^{n-k-1}\binom{n-k-p+1}{2}\binom{p+k-2}{p} = \binom{n}{k-1}$, the first line in (6.6) becomes $D(k)D(n-k)\binom{n}{k-1}$. In a similar vein, the second line of (6.6) can be re-written as $D(n-k)D(k)\binom{n}{n-k-1}$. This completes our proof of the proposition. ∎

Acknowledgments
The author gratefully acknowledges Jotun Hein and Mike Steel for useful comments on the manuscript. This research is supported by EPSRC under grant HAMJW, by MRC under grant HAMKA, and by a grant from the Danish Natural Science Foundation (SNF-5503-13370). We acknowledge Oxford Supercomputing Centre for allowing us to use their CPU time.

References

1. B.L. Allen and M. Steel, *Subtree Transfer Operations and Their Induced Metrics on Evolutionary Trees,* Ann. Comb. **5** (2001) 1-13.

2. J. Hein, *A Heuristic Method to Reconstruct the History of Sequences Subject to Recombination,* J. Mol. Evol. **36** (1993) 396-405.

3. E. Schröder, *Vier Combinatorische Probleme,* Zeit. für. Math. Phys. **15** (1870) 361-376.

4. Y.S. Song and J. Hein, *Parsimonious Reconstruction of Sequence Evolution and Haplotype Blocks: Finding the Minimum Number of Recombination Events,* to appear in: *Lecture Notes in Computer Science, Proceedings of Workshop on Algorithms in Bioinformatics 2003,* Springer Verlag.

5. D.L. Swofford and G.J. Olsen, *Phylogeny Reconstruction,* in: *Molecular Systematics,* D.M. Hillis et al., Eds., Sinauer Associates, Massachusetts, 1990, pp. 411-501.