Tanglegrams: A Reduction Tool for Mathematical Phylogenetics

Frederick A. Matsen IV, Sara C. Billey, Arnold Kas, and Matjaž Konvalinka

Abstract—Many discrete mathematics problems in phylogenetics are defined in terms of the relative labeling of pairs of leaf-labeled trees. These relative labelings are naturally formalized as tanglegrams, which have previously been an object of study in coevolutionary analysis. Although there has been considerable work on planar drawings of tanglegrams, they have not been fully explored as combinatorial objects until recently. In this paper, we describe how many discrete mathematical questions on trees "factor" through a problem on tanglegrams, and how understanding that factoring can simplify analysis. Depending on the problem, it may be useful to consider a unordered version of tanglegrams, and/or their unrooted counterparts. For all of these definitions, we show how the isomorphism types of tanglegrams can be understood in terms of double cosets of the symmetric group, and we investigate their automorphisms. Understanding tanglegrams better will isolate the distinct problems on leaf-labeled pairs of trees and reveal natural symmetries of spaces associated with such problems.

Index Terms—Phylogenetics, combinatorics, abstract algebra

1 Introduction

Consider the problem of computing the *subtree-prune-regraft* (SPR) distance between two leaf-labeled phylogenetic trees. An SPR move cuts one edge of the tree and then reattaches the resulting rooted subtree at another edge (Fig. 1). The SPR distance between two (phylogenetic, meaning leaf-labeled) trees T_1 and T_2 is the minimum number of SPR moves required to transform T_1 into T_2. This distance is of fundamental importance in phylogenetics, and many papers have been written both applying [1], [2] and investigating properties of [3], [4], [5] this distance.

Say that we wanted to calculate the SPR distance between every pair of trees on a certain number of leaves. Naïvely this would require a large number of SPR calculations, namely the number of leaf-labeled phylogenetic trees choose two. However, the distance between two such trees does not depend on the actual labels of T_1 and T_2, so one can permute the leaf labels without changing the distance. Furthermore, a path made by intermediate trees between the two trees could also have its labels permuted in order to give a path between the trees with permuted leaf labels. Thus, problems like SPR distance do not concern the actual leaf labels as such, but rather use the leaf labels as markers that can be used to map leaves of one phylogenetic tree on to another: the problem and its solutions are actually defined in terms of a *relative* leaf labeling (Fig. 1).

Analogous discrete mathematics problems and objects defined in terms of tuples of labeled combinatorial objects, but without direct reference to the labels themselves, are ubiquitous in computational biology. Any distance between pairs of trees that is computed in terms of tree modifications, such as (rooted or unrooted) subtree-prune-regraft described above, *nearest-neighbor-interchange* and *tree bisection and reattachment* (see [4] for a review), satisfy this condition. Such moves are used as the basis of both maximum-likelihood heuristic search and Bayesian Markov chain Monte Carlo (MCMC) tree reconstruction. The corresponding graph, in which trees form vertices and a collection of moves form edges, has natural symmetries of pairs of points in these spaces which have the same relative labeling. For example, hitting times of simple random walks on graphs formed by such moves for given start and end trees [6], [7], [8] are defined in terms of relative labelings between the start and end trees. The same is true for more complex random walks such as Markov chain Monte Carlo using a label-invariant likelihood, as would be used for sampling from a prior distribution on trees [9]. Graph characteristics such as Ricci-Ollivier curvature [10] under simple random walks or MCMC with a label-invariant likelihood are expressed in terms of relative tree labelings [11]. Analogous considerations hold for the problem of species delimitation, which can naturally be phrased in terms of inference of a partition of relatively labeled objects: neither distances between partitions [12] nor the graphs underlying MCMC over these partitions [13] actually refer to labels themselves.

The concept of a pair of rooted phylogenetic trees with a relative leaf labeling has been formalized as a *tanglegram* [14], [15], [16], [17]. A tanglegram is a pair of trees on the same set of leaves with a bijection between the leaves in the two trees [18] (Fig. 2). Any of the problems described above in terms of a relative labeling of a pair of phylogenetic trees can thus be expressed as a problem on tanglegrams. Thus, we can say that these problems in the discrete mathematics of phylogenetic trees "factor" through a problem on tanglegrams. The map from pairs of trees to tanglegrams is a many-to-one mapping, resulting in a substantial computational reduction for situations in which one would like to solve such a problem on many or all pairs of trees. There has been extensive work on the problem of finding the layout of a given tanglegram in the plane that minimizes crossings, with the goal of most clearly visualizing co-evolutionary relationships between species [18], [19], [20], [21], [22], [23].

However, we are not aware of any work considering tanglegrams as a convenient formalization of the notion of a relative leaf labeling in the context of studying pairs of labeled phylogenetic trees. There has also been little work enumerating or finding other properties of tanglegrams until a recent burst of activity stimulated by our work [24], [25], [26], [27], [28]. In addition, other challenging and important problems in mathematical phylogenetics reduce to questions on relatively-labeled collections of more than two trees, and correspondingly one can extend the notion of tanglegram to more than two trees. For example, "supertree" methods reconstruct a tree from collections of trees, each of which is typically considered to express information about the larger tree [2], [29], [30], which in fact is a problem on multi-tree tanglegrams. The same is true for the minimal hybridization network [31] and maximum agreement subtree [32], [33] problems. Thus even more problems in the discrete mathematics of phylogenetic trees factor through a problem concerning a multi-tree version of a tanglegram in the sense described above, which are called *tangled chains* below.

With this motivation for studying tanglegrams in more depth, in this short paper we formalize more general notions of tanglegram, describe their symmetries, observe that tanglegrams have a convenient algebraic formulation as double cosets of the symmetric group, provide some enumeration results for four types of tanglegram, and provide an introduction to the latest work on tanglegrams.

2 Tanglegrams

An *unrooted binary tree* T is a finite graph for which there is a unique path between every pair of vertices, and such that every non-leaf vertex has degree three. A *rooted tree* is an unrooted tree with a distinguished node called the *root*. We will make the

- *F.A. Matsen and A. Kas are with the Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109. E-mail: ematsen@gmail.com, akas@uw.edu.*
- *S.C. Billey is with the Department of Mathematics, University of Washington, Seattle, WA 98195. E-mail: billey@math.washington.edu.*
- *M. Konvalinka is with the Department of Mathematics, University of Ljubljana, Ljubljana 1000, Slovenia. E-mail: matjaz.konvalinka@gmail.com.*

Manuscript received 16 July 2015; revised 27 Aug. 2016; accepted 21 Sept. 2016. Date of publication 3 Oct. 2016; date of current version 2 Feb. 2018.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2016.2613040

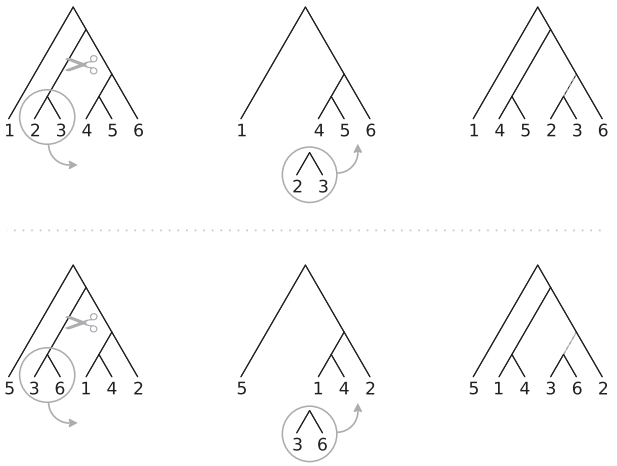

Fig. 1. Two equivalent subtree-prune-regraft moves applied to trees which are identical up to relabeling. The number of such moves required to transform one tree into another only depends on the *relative* leaf labeling between the two trees.

assumption common in phylogenetics that the root of a rooted tree has degree two, and that there are no degree-two nodes other than the root (if there is a root). The *leaves* $L(T)$ of a tree T are degree-one vertices of the tree.

Definition 1. *Let T and S be trees with the same number of leaves. An* ordered tanglegram *Y on (T, S) is an ordered triple (T, ϕ, S), where ϕ is a bijection $L(T) \to L(S)$.*

The *graph of the tanglegram* Y is the graph formed from the union of T and S by adding an edge from each leaf x in T to the corresponding leaf $\phi(x)$ in S. We will distinguish these *between-leaf edges* from the *tree edges* of T and S (Fig. 2).

We have defined tanglegrams in terms of ordered triples $Y = (T, \phi, S)$, so $Y' = (S, \phi^{-1}, T)$ is a different tanglegram. This is a sensible definition when considering sequences of trees with an inherent directionality. However, often there is not such a directionality, such as for subtree-prune-regraft moves, which are easily reversed. This motivates the following concept:

Definition 2. *A* unordered tanglegram *is a pair $(\{T, S\}, \phi)$ where $\{T, S\}$ is an unordered set of two trees, and ϕ is a bijection between $L(T)$ and $L(S)$.*

2.1 Automorphisms and Tanglegram Equivalence

Let $V(X)$ denote the vertex set of a graph X. An *isomorphism* between unrooted trees T and S is a bijective map $h : V(T) \to V(S)$ in which f maps edges of T to edges of S. For a rooted tree, we add the requirement that an isomorphism must map the root node of T to the root node of S. An *automorphism* of a tree T is an isomorphism of T with itself. It is clear that the degree of each node (i.e., the number of adjacent nodes) is preserved under isomorphisms. In phylogenetics, it is common that the root of a tree is the only node of degree two. In this case, there is no distinction between isomorphisms of rooted trees and isomorphisms of these trees as unrooted trees because degrees are preserved under isomorphism.

We start with an "obvious" lemma. First note that any isomorphism between trees T and S preserves the leaf sets $L(T)$ and $L(S)$, and therefore induces a bijection between $L(T)$ and $L(S)$.

Lemma 3. *An isomorphism between (rooted or unrooted) trees T and S is uniquely determined by the induced bijection between $L(T)$ and $L(S)$. In particular, an automorphism of a tree T is uniquely determined by the induced permutation of the leaf set $L(T)$.*

Thus we will often use tree isomorphisms $T \to S$ and bijections $L(T) \to L(S)$ interchangeably. We can now define the notion of isomorphism for tanglegrams.

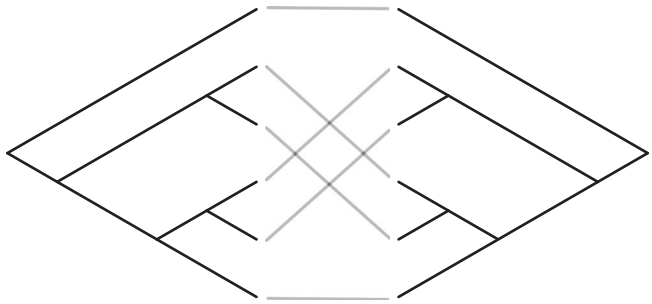

Fig. 2. The tanglegram corresponding to the pairs of trees in Fig. 1, with the bijection shown in gray. Any pair of trees with leaf labels matching as indicated will have the same subtree-prune-regraft (or any other) distance between them. When considered as a single graph, the black edges are called *tree edges*, and the gray edges are called *between-leaf edges*.

Definition 4. *Given two tanglegrams $Y = (T, \phi, S)$ and $Y' = (T, \phi', S)$ on the same pair of trees, an isomorphism of Y and Y' is defined by a pair of automorphisms $g : L(T) \to L(T)$ and $h : L(S) \to L(S)$ satisfying $h \circ \phi = \phi' \circ g$.*

The condition in the definition can be visualized in the commutative diagram

$$\begin{array}{ccc} L(T) & \xrightarrow{\phi} & L(S) \\ \downarrow{\scriptstyle g} & & \downarrow{\scriptstyle h} \\ L(T) & \xrightarrow{\phi'} & L(S). \end{array}$$

Note that if two tanglegrams Y_1 and Y_2 are isomorphic, then there is a 1-1 map from the graph of Y_1 to the graph of Y_2 which maps between-leaf edges to between-leaf edges.

2.2 Symmetries of Trees

In order to describe the ensemble of tanglegrams it is necessary to review the symmetries of the trees in the tanglegram. Although this material is classical, we were not able to find a simple presentation, and so provide one here. We will assume familiarity with the basics of group theory (covered by dozens of textbooks, e.g., [34]). Automorphisms of a tree T form a group under composition. Using $\mathfrak{S}_n$ to denote the symmetric group on n objects, leaf automorphisms of T form a subgroup $A(T)$ of $\mathfrak{S}_{|L(T)|}$.

To enumerate symmetries of trees it is convenient to use the notion of a *wreath product*; we will only define and use wreath product in the case when the acting group is $\mathfrak{S}_k$. Use G^k to denote the k-fold direct product $G \times \cdots \times G$, which has group structure given by component-wise application of G's group operation.

Given a group G, the wreath product $G \wr \mathfrak{S}_k$ of G by $\mathfrak{S}_k$ can be described as the direct product $G^k \times \mathfrak{S}_k$ with the following group operation. An element of $\mathfrak{S}_k$ acts on G^k by permuting the components, such that the group action of $\sigma \in \mathfrak{S}_k$ on $g \in G^k$ is the element $\sigma(g) \in G^k$ with ith component $g_{\sigma(i)}$. Given elements g, g' in G^k and $\sigma, \sigma' \in \mathfrak{S}_k$, the wreath group law is

$$(g, \sigma)\,(g', \sigma') := (g\,\sigma(g'), \sigma\sigma').$$

For rooted trees, Jordan [35] and Pólya [36] observed that the automorphism group of any rooted tree can be built by repeated direct products and wreath products of symmetric groups as follows. In the simplest case, assume a rooted tree T for which the root has two children subtrees T_1 and T_2. If T_1 and T_2 are isomorphic (and thus have the same automorphism groups), the automorphism group of T is the wreath product $A(T_1) \wr \mathfrak{S}_2$. That is, its symmetry group is two copies of $A(T_1)$ along with the symmetry exchanging T_1 and T_2, equipped with the group operation that

appropriately exchanges the subtrees before applying symmetries to the subtrees. If T_1 and T_2 are not isomorphic, then $A(T)$ is simply the direct product $A(T_1) \times A(T_2)$.

Now let T be a tree whose root has some number of children, each of which are roots of subtrees $T_1, \ldots, T_r$. We can reorder and partition the subtrees into N sets:

$$\{T_1, \ldots, T_{i_1}\}, \{T_{i_1+1}, \ldots, T_{i_2}\}, \ldots, \{T_{i_{N-1}+1}, \ldots, T_{i_N}\},$$

such that the subtrees in each set of the partition are isomorphic to one another and the subtrees in different sets are not isomorphic. This defines integers $i_1, \ldots, i_N$; take i_0 to be zero. A more general version of the argument above establishes

Theorem 5 (Jordan, 1869). *$A(T)$ is the direct product $A_1 \times \cdots \times A_N$, where A_j is the wreath product of $A(T_{i_j})$ with the symmetric group $\mathfrak{S}_{i_j - i_{j-1}}$.*

This defines the automorphism group of a rooted tree recursively, where of course the automorphism group of a single leaf is trivial.

Example 6. Let T_n denote the perfectly balanced binary tree on 2^n leaves and let $G_n = A(T_n)$. $G_2 = \mathfrak{S}_2$ and for each n, $G_n = G_{n-1} \wr \mathfrak{S}_2$. Moreover, $|G_n| = 2|G_{n-1}|^2$.

Example 7. The symmetry group of the Newick-format [37] tree $(1, ((2, 3), ((4, 5), 6)))$; (shown as the upper-left tree of Fig. 1) is the direct product of the symmetry groups of $(2, 3)$ and $((4, 5), 6)$. Each of these symmetry groups is $\mathfrak{S}_2$.

The automorphism group of an unrooted tree will become clear after we describe a classical and mathematically natural way to root an unrooted tree: at the *centroid*. Let T be a tree, and let x be a node of T. If we remove x as well as the edges attached to x from T, we obtain a number of disjoint connected and rooted subtrees, $X_1, \ldots, X_k$.

Definition 8. *The weight of x, $w(x)$, is defined as the maximum number of nodes of the subtrees $X_1, \ldots, X_k$.*

Definition 9. *The node x is said to be a centroid of T if $w(x)$ is minimal over all nodes of T.*

It is clear that any automorphism of T maps a centroid to a centroid, a fact which we will use to find a root fixed under leaf automorphism. Centroids are unique or nearly so, as shown by the following theorem, the proof of which can be found as a guided exercise in [38, Section 2.3.4.4].

Theorem 10 (Jordan, 1869). *Every tree has either:*

1) *a unique centroid or*
2) *two adjacent centroids.*

In case 2, every automorphism either preserves the centroids or exchanges them.

Let T be an unrooted tree, and let T_r be the rooted tree formed by rooting T at either the unique centroid, or at a new node in the edge joining a pair of centroids.

Corollary 11. *The automorphism group of an unrooted tree T is identical to the automorphism group of the associated rooted tree T_r.*

Example 12. The symmetry group of the six-leaf unrooted tree with three two-leaf subtrees (Newick format $((1, 2), (3, 4), (5, 6));$) is $\mathfrak{S}_2 \wr \mathfrak{S}_3$.

2.3 Double Cosets and Enumeration of Tanglegrams

We are now ready to algebraically describe the set of tanglegrams on a pair of n-leaf trees. Assume n-leaf trees T and S, which are both rooted or both unrooted. Arbitrarily mark the elements of the leaf sets $L(T)$ and $L(S)$ with the same set of n labels, such that we can identify both $A(T)$ and $A(S)$ as subgroups of $\mathfrak{S}_n$. Using this same marking, we can also think of the bijections from $L(T)$ to $L(S)$ as being elements of $\mathfrak{S}_n$, thus these elements of $\mathfrak{S}_n$ define tanglegrams on T and S. Recall Definition 4, stating that the set of bijections ϕ' giving the same tanglegram as a given ϕ are those for which there exist automorphisms $g \in A(T)$ and $h \in A(S)$ such that $h \circ \phi = \phi' \circ g$. This criterion is equivalent to $\phi' = h\phi g^{-1}$ as group elements in $\mathfrak{S}_n$. The set of elements satisfying such a criterion is called a double coset [34].

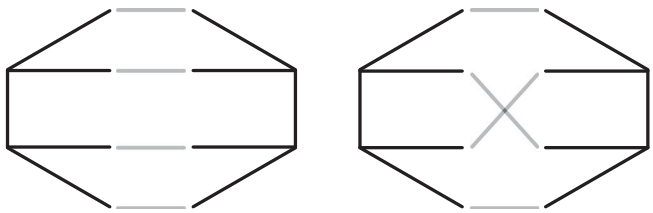

Fig. 3. The two unrooted binary tanglegrams with four leaves.

Definition 13. *Given a subgroup J of a group G and $g \in G$, the* right coset *Jg (resp.* left coset *gJ) of J in G is the set of elements of the form $\{jg \mid j \in J\}$ (resp. $\{gj \mid j \in J\}$). The number of right cosets of J in G is equal to the number of left cosets. This number is defined as the index of J in G and is denoted $[G : J]$. Given two subgroups J and K of G, the* double coset *JgK for some $g \in G$ is the set of elements $\{jgk \mid j \in J, k \in K\}$.*

Any two right (left) cosets of J in G are either identical or disjoint, and the number of elements in any coset is the same, i.e., $|J|$. In contrast to single cosets (left or right), the number of elements in a double coset may vary. We state these observations, and the equivalent observations in the unordered case, as a proposition.

Proposition 14. *Given two trees T and S with n leaves,*

- *the set of tanglegrams isomorphic to an ordered tanglegram (T, w, S) is in 1-1 correspondence with the double coset $A(S)wA(T)$ of $\mathfrak{S}_n$.*
- *the set of unordered tanglegrams isomorphic to $(\{T, S\}, w)$ is in 1-1 correspondence with equivalence classes of double cosets $A(S)wA(T)$ where pairs of cosets HwK and $Kw^{-1}H$ are deemed equivalent.*

Note that the actual 1-1 correspondence depends on the marking of T and S. Some useful facts on cosets [34], [39]:

- any two cosets are either disjoint or identical
- every double coset is a disjoint union of right cosets and a disjoint union of left cosets
- the number of right cosets of H in HgK is the index $[K : K \cap g^{-1}Hg]$, and the number of left cosets of K in HgK is the index $[H : H \cap gKg^{-1}]$.

Combining these facts with the proposition above:

Proposition 15. *The number of bijections from $L(T)$ to $L(S)$ giving an ordered tanglegram isomorphic to $Y = (T, w, S)$ is equal to $|A(S)|[A(T) : A(T) \cap w^{-1}A(S)w]$, or equivalently $|A(T)|[A(S) : A(S) \cap wA(T)w^{-1}]$.*

Example 16. Let T and S be the unique binary unrooted tree with four leaves. There are two distinct tanglegrams on (T, S) in both the ordered and unordered cases (Fig. 3). The automorphism group of either tree, $A(T)$, is the wreath product of $\mathfrak{S}_2$ by $\mathfrak{S}_2$, thus of order 8 (set theoretically $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$). Marking the leaves with the integers 1 through 4 such that $(1, 2)$ and $(3, 4)$ are both sister pairs, $G = A(T)$ is generated by $\{(12), (34), (13)(24)\} \subset \mathfrak{S}_4$.

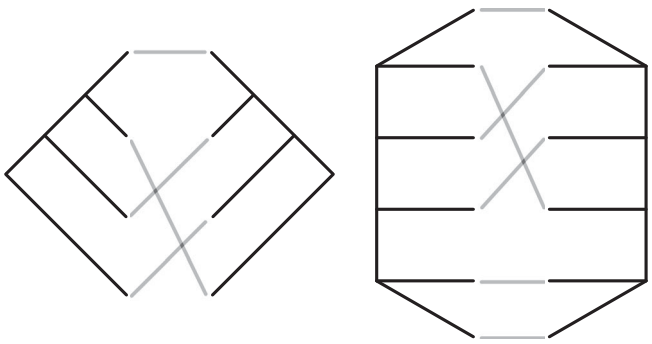

Fig. 4. Unordered rooted and unrooted tanglegrams formed by two copies of the same tree but with no automorphism that switches the trees forming each tanglegram. These examples show that the second condition of Proposition 18 is not always satisfied.

The symmetric group $\mathfrak{S}_4$ contains $4! = 24$ elements. Every double coset is a disjoint union of single cosets, and G contains eight elements, therefore the number of elements in a double coset is a multiple of 8. Moreover, since the double cosets partition $\mathfrak{S}_4$, we either have three double cosets (each of eight elements), or two double cosets (one of eight elements and one of 16 elements), or one coset (of 24 elements). Taking $w = (23)$, we calculate:

$$G \cap wGw^{-1} = \{(), (12)(34), (13)(24), (14)(23)\}.$$

Using the properties of double cosets, we find that the number of single cosets in the double coset GwG is the index $[G : G \cap w^{-1}Gw] = 2$. Thus this double coset has 16 elements, and so there must be two double cosets, corresponding to the two tanglegrams.

2.4 Symmetries of Tanglegrams

Definition 17. *An automorphism of an ordered tanglegram Y is an automorphism of the graph of Y which maps each tree to itself. An automorphism of an unordered tanglegram Y is an automorphism of the graph of Y which preserves the between-leaf edges, so an automorphism of an unordered tanglegram either maps each tree to itself or switches the two trees. If Y is a rooted tanglegram, then an automorphism of Y is required to preserve the roots of the two trees.*

If the automorphism $f : Y \to Y$ exchanges the two trees, f is described by a pair of isomorphisms: $g_1 : T \to S$ and $g_2 : S \to T$. For any leaf x of T, the image of a bijective pair $(x, \phi(x))$ must map to another bijective pair $(g_2(\phi(x)), g_1(x))$. This implies that $g_1(x) = \phi(g_2(\phi(x)))$, and thus in general that $g_1 = \phi \circ g_2 \circ \phi$. If we put the same set of distinguishing marks on the leaves of the trees T and S, we may again consider the bijection ϕ to be an element of the symmetric group $\mathfrak{S}_n$. With these conventions, we have shown that there exist $g_1 \in A(T)$ and $g_2 \in A(T)$ such that $g_1 = \phi\, g_2\, \phi$ as group elements when there is an automorphism that switches the two trees. The converse follows from reversing this argument. In summary:

Proposition 18. *If Y is an unordered tanglegram, then there exists an automorphism of Y that switches the two trees if and only if:*

- *the trees T and S are isomorphic, and*
- *$A(T) \cap \phi A(T)\phi \neq \emptyset$.*

On the other hand, if $f : Y \to Y$ is an automorphism which maps each tree to itself, then f is described by two automorphisms $g : T \to T$ and $h : S \to S$ satisfying $\phi \circ g = h \circ \phi$ when restricted to the leaves, or $g = \phi^{-1}h\phi$ as elements of the symmetric group.

Proposition 19. *Assume an ordered tanglegram $Y = (T, \phi, S)$, or an unordered tanglegram $(\{T, S\}, \phi)$. Set $H = A(T) \cap \phi^{-1}A(T)\phi$.*

1) *If Y is ordered or T is not isomorphic to S: $A(Y) = H$.*
2) *If Y is unordered and T is isomorphic to S:*
   a) *if $A(T) \cap \phi A(T)\phi \neq \emptyset$, then $A(Y)$ contains H as a subgroup of index 2.*
   b) *otherwise, $A(Y) = H$.*

Similar to the case for trees, tanglegram automorphisms are determined entirely by their action on the leaves of one of the trees.

2.5 Labeled Tanglegrams

Analogous to the concept of a leaf-labeled tree, there is a concept of a labeled tanglegram.

Definition 20. *A* labeled tanglegram *is a tanglegram along with a bijective map of the label set X to the leaves of one of the trees.*

This is analogous to the definition of a leaf-labeled phylogenetic tree [40]. The other tree can be considered to be labeled by the composition of the labeling with the bijection. Applying this labeling to both trees and then forgetting the bijection gives a pair of leaf labeled trees on the same label set, and each such pair of leaf labeled trees obviously determines a labeled tanglegram. Thus, labeled tanglegrams are in one-to-one correspondence with pairs of leaf-labeled phylogenetic trees. If the tanglegram is ordered, then this is an ordered pair of trees, and if unordered it is unordered.

It is natural to ask how many distinct labeled n-tanglegrams have the same underlying ordered or unordered tanglegram. Each leaf has a distinct label, such that the symmetric group acts freely on these labels. By the orbit-stabilizer theorem,

Proposition 21. *The number of leaf-distinct labelings of a given n-tanglegram Y is equal to $n!/|A(Y)|$.*

This is true for ordered and unordered tanglegrams, using their respective automorphism definitions. For example, there are 12 labelings for the ordered tanglegram $(1, (2, (3, 4)))$; $(((1, 2), 3), 4)$; but only 6 when considered as an unordered tanglegram.

We can use this proposition to obtain the expected value of a function d on uniformly sampled pairs of labeled trees, but which is constant on pairs of trees that form the same tanglegram (such as SPR distance or Ricci-Ollivier curvature). Dropping the denominator of the number of pairs of trees squared,

$$\begin{aligned}\sum_{T_1,T_2} d(T_1, T_2) &= \sum_{Y} \sum_{(T_1,T_2)=Y} d(T_1, T_2) \\ &= d(Y) \sum_{Y} |\{T_1, T_2 \,|\, (T_1, T_2) = Y\}| \\ &= \sum_{Y} d(Y)\, n!/|A(Y)|,\end{aligned}$$

where we use $(T_1, T_2) = Y$ to denote that T_1 and T_2 form tanglegram Y and $d(Y)$ is the common value of d applied to any pair of trees forming a tanglegram Y.

3 Variants and Special Cases

3.1 Multiple Trees

The definition of a tanglegram on two trees can be generalized to a version on multiple trees.

Definition 22. *Given trees $T_1, \ldots, T_n$ with the same number of leaves, a* tangled chain *on this set of trees is given by a pair of tuples $((T_1, \ldots, T_n), (\phi_{ij})_{i,j\in 1,\ldots,n})$ in which $\phi_{ij} : L(T_i) \to L(T_j)$ are bijections satisfying:*

1) *$\phi_{ii} = 1$ for all i;*
2) *$\phi_{ji} = \phi_{ij}^{-1}$ for all i, j;*
3) *$\phi_{ik} = \phi_{jk} \circ \phi_{ij}$, for all i, j, k.*

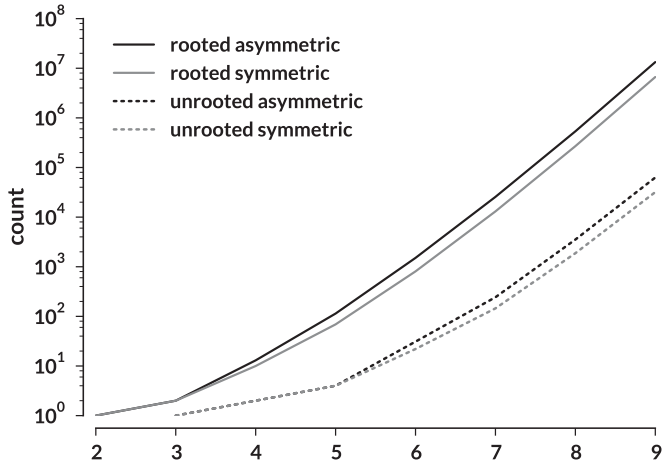

Fig. 5. Counts of various types of tanglegrams.

We can also generalize the definition of isomorphism to tangled chains on n trees.

Definition 23. *Two tangled chains on the same list of trees* $Y = ((T_1, \ldots, T_n), (\phi_{ij})_{i,j\in 1,\ldots,n})$ *and* $Y' = ((T_1, \ldots, T_n), (\phi'_{ij})_{i,j\in 1,\ldots,n})$ *are isomorphic if there exist automorphisms* $(g_i : T_i \to T_i)_{i\in 1,\ldots,n}$ *and* $(h_i : T_i \to T_i)_{i\in 1,\ldots,n}$ *satisfying* $h_j \circ \phi_{ij} = \phi'_{ij} \circ g_i$ *for* $i, j = 1, \ldots, n$.

It is clear that the n^2 bijections ϕ_{ij} are completely determined by the $n-1$ bijections $\{\phi_{1i}\}_{i=2,\ldots,n}$, since $\phi_{ij} = \phi_{1j} \circ \phi_{1i}^{-1}$. With this observation, we can rephrase the definition of isomorphism above:

Proposition 24. *Using the notation above, tangled chains* Y_1 *and* Y_2 *are isomorphic if and only if there exist automorphisms* $g_i \in A(T_i), i = 1, \ldots, n$ *satisfying* $\phi'_{1i} = g_i \circ \phi_{1i} \circ g_1^{-1}$.

Alternatively, the automorphisms ϕ_{ij} are completely determined by a sequence $\phi_{12}, \phi_{23}, \ldots, \phi_{k-1\,k}$, motivating use of the term *tangled chains*.

3.2 Partitions

Another line of inquiry in computational evolutionary biology concerns species delimitation, which can naturally be phrased in terms of inference of a partition of labeled objects. In a manner analogous to phylogenetic trees, researchers use MCMC to explore the posterior on such partitions [13], and comparison of the results can be performed using distances between the partitions [12]. Similar considerations hold for random walks and these distances as described in the introduction for trees. These partitions can also be thought of as a certain type of leaf-labeled tree of height two, thus pairs of partitions on the same underlying set also give a type of tanglegram.

All of the above conclusions hold for such partition tanglegrams as well. The automorphisms of a partition are a special case of Theorem 5. For example, the partition $123\,|\,456\,|\,78$ has automorphism group $(\mathfrak{S}_3 \wr \mathfrak{S}_2) \times \mathfrak{S}_2$.

4 Enumeration

Using a computer algebra package such as GAP4 [41] which is able to enumerate double cosets, and a package such as Sage [42] which can obtain symmetry groups of graphs, one can apply Proposition 14 to directly enumerate any type of tanglegram on a given pair of trees. We have provided code to do so at https://github.com/matsengrp/tangle, along with a script to plot tanglegrams in the plane. Although this code is not practical for the purposes of counting tanglegrams compared to the methods of [24], [25], it is very useful for work such as [11] in which one needs a complete list of tanglegrams rather than just a count. This code can work with the various types of tanglegrams (Fig. 5, Table 1).

TABLE 1
Enumeration of Binary Tanglegrams of Four Types: Rooted Ordered, Rooted Unordered, Unrooted Ordered, and Unrooted Unordered

n	rooted ord	rooted unord	unroot ord	unroot unord
1	1	1	1	1
2	1	1	1	1
3	2	2	1	1
4	13	10	2	2
5	114	69	4	4
6	1,509	807	31	22
7	25,595	13,048	243	145
8	535,753	269,221	3,532	1,875
9	13,305,590	6,660,455	62,810	31,929

More terms of these sequences can be found on the OEIS [43] as A258620, A259114, A259115, and A259116.

Although such direct enumeration procedures were state of the art when this paper was written, a number of recent papers by combinatorics experts have appeared solving various counting problems. We will now review this work. The first such advance was an elegant exact formula for the total number of binary ordered rooted tanglegrams on n leaves and for the corresponding tangled chains [24]. This work also shows that the number of (binary ordered rooted) tanglegrams is $O(n!\,4^n\,n^{-3})$. Thus there are many fewer such tanglegrams than there are pairs of leaf-labeled trees. Indeed, a simplification of the argument establishing Corollary 8 of [24] shows that the ratio of the number of ordered pairs of leaf-labeled rooted trees to the number of binary ordered rooted tanglegrams t_n is asymptotically

$$\frac{((2n-3)!!)^2}{t_n} \sim \frac{n!}{e^{1/8}}.$$

Thus by considering tanglegrams rather than pairs of labeled phylogenetic trees one obtains an asymptotically $n!$ decrease in complexity.

Next Ira Gessel [25] applied a powerful tool in combinatorics called (coincidentally) the theory of *species*. In this context, species are classes of combinatorial objects, and natural operations on those combinatorial objects correspond to category-theoretic operations on the species. For example, the fact that rooted binary trees are themselves the result of joining rooted binary trees at the root can be translated into an expression in terms of species. Having such a species expression in hand leads directly to enumerative theory. This gave further insight into the formula of [24], as well as a faster method for computing the number of unordered rooted tanglegrams and both ordered and unordered unrooted tanglegrams of small size.

Other work has followed up on the procedure for random generation of tanglegrams from [24]. Konvalinka and Wagner [26] investigate the shape of random tanglegrams, proving that the two halves of a random tanglegram look essentially like two independently chosen random plane binary trees. Éric Fusy [27] gives a more "canonical" proof of a central formula from [24] which leads to a simplification of the random tanglegram generation algorithm. Czabarka et al. [28] showed that the number of crossings (e.g., the minimal number of intersections of the gray lines in Fig. 2) of a randomly sampled tanglegram with n leaves is at least quadratic in n with high probability.

5 Discussion

Tanglegrams have been an object of study since before DNA sequences were widely available for the reconstruction of phylogenetic trees [44]. Until recently they have been studied before in the context of co-evolutionary analyses, classically that between a host and a parasite, a subject of continuing interest [45], [46]. As such, there has been extensive work on the case in which two rooted trees are distinguished between one another, as when one tree

represents hosts and one parasites, which we call the ordered rooted case. Here we have broadened the definition of tanglegrams by considering a broader class of underlying graphs, including unordered and/or unrooted tanglegrams.

In this form, tanglegrams formalize statements concerning pairs of phylogenetic trees on the same leaf set that do not directly make reference to the labels themselves. Unordered tanglegrams also do not make reference to the order of the trees.

We observe that many problems in phylogenetic combinatorics "factor" through a problem on tanglegrams. As such, we believe tanglegrams to be a worthwhile object of study in phylogenetic combinatorics, and note that they have already been crucial in an analysis of the geometry of the subtree-prune-regraft graph [11]. The work reported here provided the first enumerative methods for tanglegrams, which as described are still useful for obtaining a complete list of tanglegrams, although they have been greatly surpassed by combinatorial methods for the purpose of counting tanglegrams. These combinatorial methods have resulted in an explicit formula for the number of rooted ordered tanglegrams and improved methods for counting other types of tanglegrams, although no analogous formula is known for these other cases.

One direction for future work involves considering more general classes of graphs. For example, the tanglegram layout problem has been studied for rooted phylogenetic networks [47]. More generally, given a natural number n, one can define an *n-leaved graph* as a graph U along with n distinguished vertices $L(U)$. Given a natural number n, one could define a *generalized n-tanglegram* as a triple (U, ϕ, V), where U and V are a pair of n-leaved graphs and ϕ is a bijection between $L(U)$ and $L(V)$. If we require that n-leaved graph automorphisms preserve the leaf set $L(U)$, we can again define the leaf automorphism group $A(U)$ to be the automorphism group of U restricted to $L(U)$. If the graphs are such that any graph automorphism is determined by its action on the leaf set, then generalized tanglegrams on a given pair of n-leaved graphs U and V are in one-to-one correspondence with double cosets $A(V)wA(U)$ in $\mathfrak{S}_n$, and some of the observations given here extend to this new case. However, it's not immediately obvious which classes of graphs considered in computational bioinformatics (e.g., level-k networks [48]) satisfy this property, or more generally the extent to which current methods used for enumeration extend to the corresponding notion of tanglegram for these other structures.

Another direction for future work involves returning to the original motivation for tanglegrams, namely to study coevolving sets of trees. The recent renewed interest in tanglegrams has developed powerful new combinatorial tools for analyzing these structures, and as described in this paper these have already been useful for studies in theoretical phylogenetics. However, these tools are not yet obviously helpful for studying coevolution. For example, the current work on random sampling on tanglegrams (described above) concerns sampling uniformly from the set of tanglegrams on some number of leaves. This does not correspond to a (nontrivial) forward-time random model of coevolution. One such model of coevolution could involve a random model of speciation for a host, and then a model of host-switching for a parasite [14], [16], [44]. Another would be a model of gene-tree evolution in species trees, which are important for inferential algorithms and thus far have been counted directly using recursive formulae [49]. Such models will generate more concordant pairs of trees than uniform sampling of tanglegrams, and even a toy model with this sort of property might be interesting to analyze and helpful in the area.

Tanglegrams naturally fill a place between two previous approaches to analyzing phylogenetic trees and their distributions, which involve considering trees with and without leaf labels. For example, for two tanglegrams to be equivalent to one another, they must have the same pair of discrete tree structures, but that is not sufficient; it is sufficient for them to come from an identical pair of leaf-labeled trees, but this is not necessary. As such, questions on random tanglegrams are not answered by previous work on sampling random trees, e.g., in the work of Bona and Flajolet [50]. We have seen a burst of interest in the combinatorics of tanglegrams since this paper and [24] were posted on preprint servers, but much work remains to be done.

Acknowledgments

The authors would like to thank Steve Evans, Ira Gessel, Michael Landis, Chris Whidden, and Bianca Viray. They also thank the authors of the Sage and GAP4 software, especially Alexander Hulpke. FAM partially supported by National Science Foundation grants DMS-1223057 and CISE-1564137, SCB partially supported by National Science Foundation grant DMS-1101017, and MK supported by Research Program Z1-5434 and Research Project BI-US/14-15-026 of the Slovenian Research Agency.

References

[1] R. G. Beiko, T. J. Harlow, and M. A. Ragan, "Highways of gene sharing in prokaryotes," *Proc. Nat. Academy Sci. USA*, vol. 102, no. 40, pp. 14332–14337, 4 Oct. 2005. [Online]. Available: http://dx.doi.org/10.1073/pnas.0504068102

[2] C. Whidden, N. Zeh, and R. G. Beiko, "Supertrees based on the subtree prune-and-regraft distance," *Systematic Biol.*, vol. 63, no. 4, pp. 566–581, 2 Apr. 2014. [Online]. Available: http://dx.doi.org/10.1093/sysbio/syu023

[3] B. L. Allen and M. Steel, "Subtree transfer operations and their induced metrics on evolutionary trees," *Ann. Combinatorics*, vol. 5, no. 1, pp. 1–15, 1 Jun. 2001. [Online]. Available: http://dx.doi.org/10.1007/s00026–001-8006-8

[4] M. Bordewich and C. Semple, "On the computational complexity of the rooted subtree prune and regraft distance," *Ann. Combinatorics*, vol. 8, no. 4, pp. 409–423, 1 Jan. 2005. [Online]. Available: http://dx.doi.org/10.1007/s00026–004-0229-z

[5] C. Whidden, R. Beiko, and N. Zeh, "Fixed-Parameter algorithms for maximum agreement forests," *SIAM J. Comput.*, vol. 42, no. 4, pp. 1431–1466, 2013. [Online]. Available: http://dx.doi.org/10.1137/110845045

[6] D. J. Aldous, "Mixing time for a Markov chain on cladograms," *Combinatorics Probability Comput.*, vol. 9, no. 3, pp. 191–204, 1 May 2000. [Online]. Available: http://journals.cambridge.org/abstract_S096354830000417X

[7] P. Diaconis and S. Holmes, "Random walks on trees and matchings," *Electron. J. Probability*, vol. 7, no. 6, pp. 1–17, 2002. [Online]. Available: http://www.emis.ams.org/journals/EJP-ECP/_ejpecp/include/EJP-2002–1307eb37.pdf?id=2856&article=1307&mode=pdf

[8] S. N. Evans and A. Winter, "Subtree prune and regraft: A reversible real tree-valued Markov process," *Ann. Probability*, vol. 34, no. 3, pp. 918–961, May 2006. [Online]. Available: http://projecteuclid.org/euclid.aop/1151418488

[9] M. E. Alfaro and M. T. Holder, "The posterior and the prior in Bayesian phylogenetics," *Annu. Rev. Ecology Evolution Syst.*, vol. 37, pp. 19–42, Jan. 2006. [Online]. Available: http://dx.doi.org/10.2307/30033825

[10] Y. Ollivier, "Ricci curvature of Markov chains on metric spaces," *J. Functional Anal.*, vol. 256, no. 3, pp. 810–864, 1 Feb. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.jfa.2008.11.001

[11] C. Whidden and F. A. Matsen IV, "Ricci-Ollivier curvature of the rooted phylogenetic subtree-Prune-regraft graph," 1 Apr. 2015. [Online]. Available: http://arxiv.org/abs/1504.00304

[12] D. Gusfield, "Partition-distance: A problem and class of perfect graphs arising in clustering," *Inf. Process. Lett.*, vol. 82, no. 3, pp. 159–164, 16 May 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020019001002630

[13] Z. Yang and B. Rannala, "Bayesian species delimitation using multilocus sequence data," *Proc. Nat. Academy Sci. United States America*, vol. 107, no. 20, pp. 9264–9269, 18 May 2010. [Online]. Available: http://dx.doi.org/10.1073/pnas.0913022107

[14] R. D. M. Page, "Parasites, phylogeny and cospeciation," *Int. J. Parasitology*, vol. 23, no. 4, pp. 499–506, Jul. 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0020751993900392

[15] M. A. Charleston, "Jungles: A new solution to the host/parasite phylogeny reconciliation problem," *Math. Biosci.*, vol. 149, no. 2, pp. 191–223, May 1998. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9621683

[16] R. D. M. Page, *Tangled Trees: Phylogeny, Cospeciation, and Coevolution*. Chicago, IL, USA: Univ. Chicago Press, 2003. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=t_NpAyxmMsgC&oi=fnd&pg=PR7&ots=tWIdAnxsN9&sig=55yn2Stt-HijZNcLwO5bYDHMGeE

[17] M. A. Charleston and S. L. Perkins, "Traversing the tangle: Algorithms and applications for cophylogenetic studies," *J. Biomed. Informat.*, vol. 39, no. 1, pp. 62–71, Feb. 2006. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2005.08.006

[18] B. Venkatachalam, J. Apple, K. St John, and D. Gusfield, "Untangling tanglegrams: Comparing trees by their drawings," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 7, no. 4, pp. 588–597, Oct.-Dec. 2010. [Online]. Available: http://dx.doi.org/10.1109/TCBB.2010.57
[19] K. Buchin, et al., "Drawing (complete) binary tanglegrams: Hardness, approximation, fixed-parameter tractability," 5 Jun. 2008. [Online]. Available: http://arxiv.org/abs/0806.0920
[20] A. Lozano, R. Y. Pinter, O. Rokhlenko, G. Valiente, and M. Ziv-Ukelson, "Seeded tree alignment," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 5, no. 4, pp. 503–513, Oct.-Dec. 2008. [Online]. Available: http://dx.doi.org/10.1109/TCBB.2008.59
[21] M. S. Bansal, W.-C. Chang, O. Eulenstein, and D. Fernández-Baca, "Generalized binary tanglegrams: Algorithms and applications," in *Bioinformatics and Computational Biology*. Berlin, Germany: Springer, 1 Jan. 2009, pp. 114–125. [Online]. Available: http://link.springer.com/chapter/10.1007/978–3-642-00727-9_13
[22] S. Böcker, F. Hüffner, A. Truss, and M. Wahlström, "A faster Fixed-Parameter approach to drawing binary tanglegrams," in *Parameterized and Exact Computation*. Berlin, Germany: Springer, 1 Jan. 2009, pp. 38–49. [Online]. Available: http://link.springer.com/chapter/10.1007/978–3-642-11269-0_3
[23] H. Fernau, M. Kaufmann, and M. Poths, "Comparing trees via crossing minimization," *J. Comput. Syst. Sci.*, vol. 76, no. 7, pp. 593–608, Nov. 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022000009001032
[24] S. Billey, M. Konvalinka, and F. A. Matsen IV, "On the enumeration of tanglegrams and tangled chains," to appear in J. Combinatorial Theory Series A, 17 Jul. 2015. [Online]. Available: http://arxiv.org/abs/1507.04976
[25] I. M. Gessel, "Counting tanglegrams with species," 13 Sep. 2015. [Online]. Available: http://arxiv.org/abs/1509.03867
[26] M. Konvalinka and S. Wagner, "The shape of random tanglegrams," Adv. in Appl. Math., vol. 78, pp. 76–93, 2016, [Online]. Available: http://dx.doi.org/10.1016/j.aam.2016.04.001
[27] E. Fusy, "On symmetries in phylogenetic trees," *Electron. J. Combin.*, vol. 23, no. 3, 24 Feb. 2016. [Online]. Available: http://arxiv.org/abs/1602.07432
[28] E. Czabarka, L. A. Székely, and S. Wagner, "Inducibility in binary trees and crossings in random tanglegrams," 26 Jan. 2016. [Online]. Available: http://arxiv.org/abs/1601.07149
[29] O. R. P. Bininda-Emonds, J. L. Gittleman, and M. A. Steel, "The (super)tree of life: Procedures, problems, and prospects," *Annu. Rev. Ecology Syst.*, vol. 33, pp. 265–289, 1 Jan. 2002. [Online]. Available: http://www.jstor.org/stable/3069263
[30] M. Steel and A. Rodrigo, "Maximum likelihood supertrees," *Syst. Biol.*, vol. 57, no. 2, pp. 243–250, Apr. 2008. [Online]. Available: http://dx.doi.org/10.1080/10635150802033014
[31] M. Baroni, C. Semple, and M. Steel, "A framework for representing reticulate evolution," *Ann. Combinatorics*, vol. 8, no. 4, pp. 391–408, 2005. [Online]. Available: http://link.springer.com/article/10.1007/s00026–004-0228-0
[32] C. R. Finden and A. D. Gordon, "Obtaining common pruned trees," *J. Classification*, vol. 2, no. 1, pp. 255–276, 1985. [Online]. Available: http://link.springer.com/article/10.1007/BF01908078
[33] M. Farach, T. M. Przytycka, and M. Thorup, "On the agreement of many trees," *Inf. Process. Lett.*, vol. 55, no. 6, pp. 297–301, 29 Sep. 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/002001909500110X
[34] D. S. Dummit and R. M. Foote, *Abstract Algebra*. Hoboken, NJ, USA: Wiley, 2004.
[35] C. Jordan, "Sur les assemblages de lignes," *J. für die reine und angewandte Mathematik*, vol. 70, pp. 185–190, 1869. [Online]. Available: http://www.math.washington.edu/~mathcircle/circle/2014-15/advanced/mc-14a-w5.pdf
[36] G. Pólya, "Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen," *Acta Math.*, vol. 68, no. 1, pp. 145–254, 1937. [Online]. Available: http://link.springer.com/article/10.1007/BF02546665
[37] Wikipedia, "Newick format—Wikipedia, the free encyclopedia," 2014. [Online]. Available: http://en.wikipedia.org/wiki/Newick_format. Accessed on: Oct. 04, 2014.
[38] D. E. Knuth, *The Art of Computer Programming: Fundamental Algorithms*, 2nd ed., vol. 1. Redwood City, CA, USA: Addison Wesley Longman, 1973.
[39] S. Lang, *Algebra*. Menlo Park, CA, USA: Addison-Wesley, 1993. [Online]. Available: http://opac.inria.fr/record=b1081613
[40] C. Semple and M. Steel, *Phylogenetics*. New York, NY, USA: Oxford Univ. Press, 2003.
[41] *GAP-Groups, Algorithms, and Programming, Version 4.7.5*, "The GAP group," 2014. [Online]. Available: http://www.gap-system.org
[42] W. Stein and D. Joyner, "SAGE: System for algebra and geometry experimentation," *ACM SIGSAM Bulletin*, vol. 39, no. 2, pp. 61–64, 2005. [Online]. Avaialble: http://sagemath.org/
[43] OEIS Foundation Inc., "The on-line encyclopedia of integer sequences," 2015. [Online]. Available: http://oeis.org
[44] M. S. Hafner and S. A. Nadler, "Phylogenetic trees support the coevolution of parasites and their hosts," *Nature*, vol. 332, no. 6161, pp. 258–259, 17 Mar. 1988. [Online]. Available: http://dx.doi.org/10.1038/332258a0
[45] R. Libeskind-Hadas and M. A. Charleston, "On the computational complexity of the reticulate cophylogeny reconstruction problem," *J. Comput. Biol.*, vol. 16, no. 1, pp. 105–117, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1089/cmb.2008.0084
[46] B. Drinkwater and M. A. Charleston, "Introducing TreeCollapse: A novel greedy algorithm to solve the cophylogeny reconstruction problem," *BMC Bioinf.*, vol. 15, no. suppl 16, 8 Dec. 2014, Art. no. S14. [Online]. Available: http://dx.doi.org/10.1186/1471–2105-15-S16-S14
[47] C. Scornavacca, F. Zickmann, and D. H. Huson, "Tanglegrams for rooted phylogenetic trees and networks," *Bioinf.*, vol. 27, no. 13, pp. i248–i256, 1 Jul. 2011. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btr210
[48] C. Choy, J. Jansson, K. Sadakane, and W.-K. Sung, "Computing the maximum agreement of phylogenetic networks," *Theoretical Comput. Sci.*, vol. 335, no. 1, pp. 93–107, 20 May 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304397504008102
[49] N. A. Rosenberg and J. H. Degnan, "Coalescent histories for discordant gene trees and species trees," *Theoretical Population Biol.*, vol. 77, no. 3, pp. 145–151, May 2010. [Online]. Available: http://dx.doi.org/10.1016/j.tpb.2009.12.004
[50] M. Bóna and P. Flajolet, "Isomorphism and symmetries in random phylogenetic trees," *J. Appl. Probability*, vol. 46, no. 4, pp. 1005–1019, 1 Dec. 2009. [Online]. Available: http://dx.doi.org/10.2307/25662477