

# Mixed-up trees: the structure of phylogenetic mixtures

Frederick A. Matsen,<sup>\*</sup> Elchanan Mossel,<sup>†</sup> and Mike Steel<sup>\*</sup>

June 2, 2007

## Abstract

In this paper we apply new geometric and combinatorial methods to the study of phylogenetic mixtures. The focus of the geometric approach is to describe the geometry of phylogenetic mixture distributions for the two state random cluster model, which is a generalization of the two state symmetric (CFN) model. In particular, we show that the set of mixture distributions forms a convex polytope and we calculate its dimension; corollaries include a simple criterion for when a mixture of branch lengths on the star tree can mimic the probability distribution on splits of a resolved quartet tree. Furthermore, by computing volumes of polytopes we can clarify how “common” non-identifiable mixtures are under the CFN model. We also present a new combinatorial result which extends any identifiability result for a specific pair of trees of size six to arbitrary pairs of trees. Next we present a positive result showing identifiability

---

<sup>\*</sup>Biomathematics Research Centre, University of Canterbury, New Zealand. email: ematsen@gmail.com, m.steel@math.canterbury.ac.nz. Supported by the Allan Wilson Centre for Molecular Ecology and Evolution.

<sup>†</sup>Statistics, UC Berkeley, email: mossel@stat.berkeley.edu. Supported by a Sloan fellowship in Mathematics, NSF awards DMS 0528488 and DMS 0548249 (CAREER) and by ONR grant N0014-07-1-05-06

of rates-across-sites models. Finally, we answer a question raised in a previous paper concerning “mixed branch repulsion” on trees larger than quartet trees under the CFN model.

Keywords: phylogenetics, model identifiability, mixture model, polytope, discrete Fourier analysis

Molecular phylogenetic inference methods reconstruct evolutionary history from sequence data. Many years of research have shown that if data evolves according to a single process under certain assumptions then the underlying tree can be found given sequence data of sufficient length. For an introduction to this literature see [3] or [10].

However, it is known that molecular evolution varies according to position even within a single gene [11]. Between genes even more heterogeneity is observed [8], though it is not unusual for researchers to concatenate data from different genes for inference [9]. This poses a different challenge for theoretical phylogenetics: is it possible to reconstruct the tree from data generated by a combination of different processes?

This question is formalized as follows. The raw data for most phylogenetic inference techniques is site pattern frequency vectors, i.e. normalized counts of how often certain data patterns occur. If multiple data sets are combined, the corresponding site pattern frequency vectors are combined according to a weighted average. In statistical terminology, this is called a “mixture model.” For brevity, we will use *a mixture of trees* to mean a mixture of site pattern

frequencies obtained from trees. When all of the trees have the same (labeled) topology we will call this *a mixture of branch lengths on a tree*. Now the above question can be written “if data is generated by a mixture of trees (branch lengths) can the trees (tree) be reconstructed from the data?”

The answer to this question is certainly “not always.” In 1994 Steel et. al. [12] presented the first “non-identifiable” examples, i.e. mixtures of branch lengths on a tree such that the underlying tree cannot be inferred from the data. More recently, Štefankovič and Vigoda [13] were the first to explicitly construct such examples. Even more recently, Matsen and Steel [6] showed that a mixture of branch lengths on one tree can “mimic” (i.e. give the same expected site pattern frequencies as) an unmixed process on a tree of another topology.

This raises several questions, some of which are answered in this paper for the two state symmetric (CFN) models and some generalizations. First, now that we know that these non-identifiable examples exist, is there some way of describing exactly which site pattern frequency vectors correspond to non-identifiable mixtures? Below we note that the set of mixture distributions on a tree of a given topology forms a convex polytope with an easy description (Proposition 5); thus the non-identifiable patterns also form a convex polytope. Now, computing dimensions shows that a “random” site pattern frequency vectors has a non-zero probability of being non-identifiable, which raises the question of the relative volumes of a given tree polytope and the non-identifiable polytopes. This question is answered by computer calculations for the quartet case in Ta-

ble 1. We also show that surprisingly well-resolved trees sit inside the mixture polytope for the star tree (Proposition 17). This same proposition implies that the internal edge of a quartet tree must be long compared to the pendant edges if the corresponding site pattern frequency vector is to be identifiable.

The second main section focuses on identifiability results for mixtures of two trees under various assumptions. These results partially “bookend” the non-identifiability results of [6, 13]. The first emphasis for this work is combinatorial, answering the question (Theorem 18) “if we know all of the splits associated to the restriction of a pair of trees to taxon subsets of size  $k$ , is it possible to reconstruct the pair of trees?” This gives a theorem which extends any identifiability result for a specific pair of trees of size six to arbitrary pairs of trees under a molecular clock. (Theorem 23). A different approach shows identifiability of rates-across-sites models for pairs of trees (Theorem 25). Finally, we show that if a mixture of two branch lengths on a single tree mimics the expected site pattern frequencies of a tree on another topology then the two topologies can differ by at most one nearest neighbor interchange.

## 1 Geometry of unbounded mixtures

In this section we show that the space of mixtures under the random cluster model is the convex hull of a finite set of points, i.e. a convex polytope. In some ways this is a surprising fact— there is no *a priori* reason why the convex hull of subvariety must be the convex hull of a finite set of points. The description

of the vertices of the polytope has some interesting consequences discussed in Section 1.2. We then compute dimensions, which is motivated in part by the following theorem due to Carathéodory:

**Theorem 1.** *If  $X$  is a  $d$ -dimensional linear space over the real numbers, and  $A$  is a subset of  $X$ , then every point of the convex hull of  $A$  can be expressed as a convex combination of not more than  $d + 1$  points of  $A$ .*

A proof can be found as statement 2.3.5 of [5]. Therefore if we know that the dimension of a certain set of phylogenetic mixture distributions is  $d$ , then any mixture distribution in that set can be expressed as a convex combination of expected site pattern frequencies from no more than  $d + 1$  trees.

We also show that the dimension of site pattern frequency vectors which can be written as mixtures of branch lengths on the star tree is equal to the corresponding dimension for all topologies together. This forms an interesting contrast to the genericity results in [1].

Convex polytopes are typically specified in one of two ways: by a *V-description*, as the convex hull of a finite set of points, or by an *H-description*, as the bounded intersection of finitely many half-spaces. Algorithms for going between the two descriptions are well-known and implemented in the software `polymake` [4]. We will make use of both descriptions; for example, the intersection of polytopes can be easily computed by taking the union of the two sets of inequalities describing the half-spaces of the *H*-descriptions. More introductory material about polytopes can be found in the texts of Grünbaum [5] and Ziegler [14].

From the phylogenetic perspective, we are interested in the set of site pattern

frequency vectors which correspond to non-identifiable mixtures. In particular, one might ask the question of which site-pattern frequency vectors can be expressed as a mixture of branch lengths on any one of a collection of tree topologies. At least in the case of the random cluster model, the answer is the intersection of the corresponding mixture polytopes. Using `polymake` and Proposition 5 this becomes an easy exercise: simply take the union of the  $H$ -description inequalities for the polytope associated with each topology. Although we focus on quartet trees in Section 1.2 similar calculations for larger trees would not be difficult.

### 1.1 The random cluster model

First we define the random cluster model, which generalizes the two state symmetric (CFN) and Jukes-Cantor DNA models [3]. For a tree  $T = (V, E)$  such models are parameterized by a distribution  $\pi$  on the  $q$  states  $[q]$  and a function  $p : E \rightarrow [0, 1]$ . The random cluster model is defined as follows:

- For each edge  $e$  declare the edge open with probability  $p(e)$ , and declare it closed otherwise. Let  $C_1, \dots, C_r$  denote the maximal closed-edge connected components of  $V$ .
- Choose  $r$  independent samples  $a_1, \dots, a_r$  from the distribution  $\pi$  and assign the state  $a_i$  to all the vertices of  $C_i$  for each  $i$ .

It is always assumed that  $\pi[x] > 0$  for all  $x \in [q]$ . We will also consider the case  $q = \infty$  in which different clusters will always be assigned different states. One

may think of the case  $q = \infty$  as a case where each state assignment to leaves partitions the set.

The CFN and Jukes-Cantor DNA models are random cluster models with  $\pi$  the uniform distribution on 2 and 4 states respectively (see, e.g., [10] p.197).

**Definition 2.** *Given a phylogenetic tree  $T = (V, E)$  with a random cluster model defined by  $p$  let  $D_{T,p}$  denote the induced distribution of state assignments to the leaves.*

**Proposition 3.** *For any tree  $T$  and any  $p$ , the distribution  $D_{T,p}$  is a convex combination of distributions  $D_{T,p_i}$  where  $p_i$  obtains only the values 0 or 1.*

*Proof.* First note that if  $p = \alpha p_1 + (1 - \alpha)p_2$  then  $D_{T,p} = \alpha D_{T,p_1} + (1 - \alpha)D_{T,p_2}$ . The claim now follows from the fact that the extremal points of  $[0, 1]^{|E|}$  are vectors all of whose coordinates are 0's and 1's. □

We begin by studying the convex structure of mixtures. Unlike the linear structure, the convex structure does not depend on the number of states  $q$ .

**Definition 4.** *Let  $\mathcal{S}$  be a partition of  $[n]$ . We denote by  $D_{\mathcal{S}}$  the distribution on  $[q]^n$  that satisfies:*

- *The random variables  $(x_S : S \in \mathcal{S})$  are independent.*
- *For all  $S \in \mathcal{S}$  and  $i, j \in S$  it holds that  $x_i = x_j$ .*
- *The marginal distributions of all the  $x_i$  are given by  $\pi$ .*

Clearly there is a unique distribution that satisfies these properties. It is easy to see that

**Proposition 5.** *Let  $T$  be a phylogenetic tree and let  $p$  be edge probabilities all of whose values are in  $\{0, 1\}$ . Then  $D_{T,p} = D_{\mathcal{S}}$  for some partition  $\mathcal{S}$  of  $[n]$ . On the other hand, for every partition  $\mathcal{S}$  of  $[n]$  there exists a phylogenetic tree  $T$  and edge probabilities  $p \in \{0, 1\}$  such that  $D_{T,p} = D_{\mathcal{S}}$ .*

In fact, the distributions  $D_{\mathcal{S}}$  determine the convex geometry of mixtures. We use  $D[f]$  to mean the expectation of  $f$  under the distribution  $D$ .

**Theorem 6.** *The set of mixture distributions on trees over  $n$  leaves is a convex set whose extremal points are given by*

$$\{D_{\mathcal{S}} : \mathcal{S} \text{ a partition of } [n]\}.$$

*Proof.* The set is clearly convex. By Propositions 3 and 5 it follows that every mixture of trees can be written as a convex sum of the elements  $D_{\mathcal{S}}$ . It thus remains to show that we cannot write  $D_{\mathcal{S}}$  as a convex combination of  $D_{\mathcal{S}_1}, \dots, D_{\mathcal{S}_k}$  if  $\mathcal{S} \notin \{\mathcal{S}_1, \dots, \mathcal{S}_k\}$ .

Assume by contradiction that

$$D_{\mathcal{S}} = \sum_i \alpha_i D_{\mathcal{S}_i},$$

where  $\alpha_i > 0$  for all  $i$  and  $\sum_i \alpha_i = 1$ .

**Claim 7.**  *$\mathcal{S}$  is a refinement of  $\mathcal{S}_i$  for all  $i$ .*

*Proof.* Suppose  $\mathcal{S}$  does not refine  $\mathcal{S}_1$ . Thus there exist  $i \neq j$  such that  $i$  and  $j$  belong to the same set in  $\mathcal{S}$  but do not belong to the same set in  $\mathcal{S}_1$ . But this

implies by definition that for  $D_{\mathcal{S}}$  we have that  $x_i = x_j$  with probability one while for  $D_{\mathcal{S}_i}$  the variables  $x_i$  and  $x_j$  are independent. This is a contradiction.  $\square$

The following claim concludes the proof of the theorem.

**Claim 8.**  $D_{\mathcal{S}}$  cannot be written as a convex combination of the  $D_{\mathcal{S}_i}$ .

*Proof.* By the previous claim, we may assume that

$$D_{\mathcal{S}} = \sum_i \alpha_i D_{\mathcal{S}_i}, \quad (1)$$

where  $\mathcal{S}$  is a refinement of each of the  $\mathcal{S}_i$ . Let

$$f(x_1, \dots, x_n) = \sum_{i,j} 1(x_i = x_j).$$

Note that for a general partition  $\mathcal{S}'$  it holds that

$$D_{\mathcal{S}'}[f] = |\mathcal{S}'|_2^2 + (n^2 - |\mathcal{S}'|_2^2)|\pi|_2^2$$

where  $|\mathcal{S}'|_2^2 = \sum_{S \in \mathcal{S}'} |S|^2$  and  $|\pi|_2^2 = \sum_{x \in [q]} \pi[x]^2$ . In particular, it follows that since  $\mathcal{S}$  is a refinement of  $\mathcal{S}_i$  and  $\mathcal{S} \neq \mathcal{S}_i$  for all  $i$ , we have  $D_{\mathcal{S}_i}[f] > D_{\mathcal{S}}[f]$  for all  $i$ . Plugging this into (1) we obtain a contradiction. The proof follows.  $\square$

$\square$

Now we calculate dimensions. The dimension of a convex polytope is defined to be the dimension of its affine hull. Unlike the convex structure, the affine structure depends crucially on the details of the model. We do not give a general

dimension formula here – instead we will just discuss the two state and infinite state models. We let  $\mathcal{D}_n(1/2, 1/2)$  denote the space of all distributions that can be written as a convex combination of phylogenetic trees on  $n$  leaves under the CFN model, and let  $\mathcal{D}_n^*(1/2, 1/2)$  denote those which can be written using sets of edge lengths on the star tree with  $n$  leaves.

**Proposition 9.**

$$\dim(\mathcal{D}_n^*(1/2, 1/2)) = \dim(\mathcal{D}_n(1/2, 1/2)) = 2^{n-1} - 1.$$

*Proof.* We will work with the Fourier transform  $F$  as follows: take the state space to be  $\{-1, 1\}$  and

$$F_S(D) = D \left[ \prod_{i \in S} x_i \right].$$

This version of the discrete Fourier transform is simply a nonzero scalar multiple of the Fourier transform defined by a Hadamard matrix ([10]) of an appropriate dimension; thus it is invertible. Since the Fourier transform is linear and invertible, and we can compute the dimension of the  $\mathcal{D}$ 's by computing the dimension of their image under the Fourier transform.

Clearly it holds that

$$F_\emptyset[D_{T,p}] = 1, \tag{2}$$

and it is known that if  $S$  is of odd size then

$$F_S[D_{T,p}] = 0, \tag{3}$$

for all  $T$  and  $p$ . This last fact can be seen as follows. By Proposition 3 we can assume that  $D_{T,p}$  is given by independent assignment of states (according to  $\pi$ ) to clusters  $C_1, \dots, C_k$ . Because  $S$  is odd, at least one of the  $S \cap C_j$  must have odd size, and

$$D \left[ \prod_{i \in S \cap C_j} x_i \right] = -1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 0.$$

Equation (3) now follows because the expectation of a product of independent random variables is the product of the expectations.

It thus follows that equalities (2) and (3) hold for all distributions in  $\mathcal{D}$ . This implies that

$$\dim(\mathcal{D}_n(1/2, 1/2)) \leq 2^{n-1} - 1.$$

We show next that

$$2^{n-1} - 1 \leq \dim(\mathcal{D}_n^*(1/2, 1/2)) \leq \dim(\mathcal{D}_n(1/2, 1/2)) \quad (4)$$

which will imply the proposition. The second inequality follows by containment.

Now we show the first inequality. Given a set  $S$ , consider the partition  $\rho(S)$  that has the sets  $S$  and a singleton set corresponding to each element of  $[n] \setminus S$ . This partition can be achieved on the star tree by declaring all of the edges in  $S$  to be closed with probability one and all of the other edges to be open with probability one. It is easy to see that

$$F_{S'}[D_{\rho(S)}] = 1 \text{ iff } S' \subset S \text{ and } S' \text{ is even.}$$

Thus clearly  $F_{S'}[D_{\rho(\emptyset)}]$  is zero for all  $S' \neq \emptyset$ . It follows (using the fact that  $F_{\emptyset}[D_{T,p}] = 1$  for any  $T, p$ ) that in this case affine dimension coincides with linear dimension. Therefore to show the first inequality of (4) it suffices to find for every set  $S$  of even order a linear combination of elements of  $\mathcal{D}_n^*(1/2, 1/2)$  whose Fourier coefficient at  $S$  is 1 and is 0 at all other sets. An inductive argument shows that in order to achieve this task, it suffices to show that for every even set  $S$  there exists an element of  $\mathcal{D}$  whose Fourier coefficient at every even subset of  $S$  is 1 and is zero on all other sets. This is exactly  $D_{\rho(S)}$  as described above. The proof follows.  $\square$

We now analyze the random cluster for  $q = 2$  when the distribution  $\pi$  is not uniform. Define  $\mathcal{D}_n^*(r, 1-r)$  and  $\mathcal{D}_n(r, 1-r)$  for the case of non-uniform  $\pi$  analogous to the symmetric (CFN) case.

**Proposition 10.** *Let  $0 < r < 1$  and  $r \neq 1/2$ . Then*

$$\dim(\mathcal{D}_n^*(r, 1-r)) = \dim(\mathcal{D}_n(r, 1-r)) = 2^n - n - 1.$$

*Proof.* Here we need a variant of the above-described Fourier transform – now we take the state space to be  $\{r-1, r\}$ , with  $\pi$  giving the first state with probability  $r$  and the second state with probability  $1-r$ . Again  $F$  will denote the Fourier transform so that

$$F_S(D) = D \left[ \prod_{i \in S} x_i \right].$$

The matrix representation of this transform in the usual basis is the  $n$ -fold Kronecker product of the matrix

$$\begin{pmatrix} 1 & 1 \\ r-1 & r \end{pmatrix}.$$

It follows that this transform is invertible. As before we calculate the dimension of the Fourier transform of the  $\mathcal{D}$ . Clearly

$$F_{\emptyset}[D_{T,p}] = 1,$$

and if  $S$  is a singleton then

$$F_S[D_{T,p}] = 0,$$

for all  $T$  and  $p$  by a similar argument to before. It thus follows that the equalities above hold for all distributions in the  $\mathcal{D}$ . This implies that

$$\dim(\mathcal{D}_n(r, 1-r)) \leq 2^n - n - 1.$$

As before, given a set  $S$ , consider the partition  $\rho(S)$  that has the sets  $S$  and a singleton set corresponding to each element of  $[n] \setminus S$ . It is then easy to see that

$$F_S[D_{\rho(S)}] = r(r-1)^{|S|} + (1-r)r^{|S|} = r(r-1) \left( (r-1)^{|S|-1} + r^{|S|-1} \right) \neq 0,$$

since  $0 < r < 1, r \neq 1/2$  and  $|S| > 1$ . On the other hand, if  $S'$  is not a subset of  $S$  then

$$F_{S'}[D_{\rho(S)}] = 0$$

by an argument as in the previous proof.

As before the affine dimension coincides with the linear dimension. To prove the corresponding lower bound it suffices to find for every set  $S$  of size at least two a linear combination of elements of  $\mathcal{D}_n^*(r, 1-r)$  whose Fourier coefficient at  $S$  is one and is zero at all other sets. An inductive argument using  $D_{\rho(S)}$  again concludes the proof.  $\square$

We have just seen how for the CFN model the affine dimension of the space of mixtures is much smaller than the number of extremal points. In contrast, for  $q = \infty$ , the dimension equals the number of extremal points. This follows from the following proposition.

**Proposition 11.** *For the  $q = \infty$  model, the distributions  $D_{\mathcal{S}}$  where  $\mathcal{S}$  runs over all partitions of  $[n]$  are linearly independent.*

*Proof.* There is nothing to prove as the probability space we are working in is the space of partitions of  $[n]$ .  $\square$

## 1.2 The mixture polytope for the CFN model

This section concerns mixtures under the CFN model. The connection with the previous section is as follows.

**Corollary 12.** *The set of mixture distributions under the CFN model on a given tree is a convex set whose extremal points are given (perhaps with repetition) by branch length assignments to that topology taken from the set  $\{0, \infty\}$ .*

*Proof.* A branch length of zero corresponds to an edge being open in the random cluster model with probability zero, and a branch length of infinity corresponds to an edge being open with probability one. The corollary now follows from Proposition 3. □

Before analyzing various associated polytopes, we fix some notation and remind the reader of some facts. For a given branch length  $\gamma$  we will call  $\theta = \exp(-2\gamma)$  the “fidelity” of an edge, which ranges between zero (infinite length edge) and one (zero length edge) for non-negative branch lengths. Denote splits for trees on  $n$  taxa using subsets  $A \subseteq \{1, \dots, n-1\}$  by specifying the “side” which doesn’t contain the  $n$ th taxa. Note that one could equivalently use even sized subsets of  $\{1, \dots, n\}$  via the  $f(A)$  below as in [6]. We will use  $p_A$  to denote the probability of a split  $A$  and  $q_A$  to denote the  $A$ th component of the Fourier transform of the split probabilities as in [6, 10]. We will denote the corresponding vectors by  $\underline{p}$  and  $\underline{q}$ . The Hadamard matrices will be denoted  $H$ ;  $H$  is symmetric and  $HH = 2^{n-1}I$  when  $H$  is  $n$  by  $n$ . We will denote inner product of  $v$  and  $w$  by  $\langle v, w \rangle$  and will often use the fact that  $\langle Hv, w \rangle = \langle v, Hw \rangle$ . We will take  $e_A$  to be the vector with  $A$ ’th component one and other components zero. We will also use the following lemma, from the the proof of Theorem 8.6.3 of [10].

**Lemma 13.** For any subset  $A \subseteq \{1, \dots, n-1\}$  of even order, let

$$f(A) = \begin{cases} A & \text{if } |A| \text{ is even} \\ A \cup \{n\} & \text{otherwise.} \end{cases}$$

Then

$$q_A = \prod_{e \in \mathcal{P}(T, f(A))} \theta(e) \quad (5)$$

where  $\mathcal{P}(T, f(A))$  is the unique set of edges which lie in the set of edge-disjoint paths connecting the taxa in  $f(A)$  to each other.  $\square$

We will abuse notation by taking  $Co(T_1, \dots, T_n)$  to denote the convex hull of site pattern frequencies on trees  $T_1, \dots, T_n$  of the same number of leaves.

There are four tree topologies on four taxa: the star tree  $T_\star$  and the three resolved trees on four taxa  $T_1$ ,  $T_2$ , and  $T_3$ . Thus, up to isomorphism, there are six convex polytopes of interest in this case, with inclusions as indicated:

$$Co(T_\star) \subseteq Co(T_1) \cap Co(T_2) \cap Co(T_3) \quad (6)$$

$$\subseteq Co(T_1) \cap Co(T_2) \quad (7)$$

$$\subseteq Co(T_1) \quad (8)$$

$$\subseteq Co(T_1, T_2) \quad (9)$$

$$\subseteq Co(T_1, T_2, T_3). \quad (10)$$

It will be shown below that the inclusion in (6) is an equality.

From a phylogenetic perspective, polytope (6) represents the site pattern

frequencies which can be realized as a mixture on any of the four topologies. Polytope (7) contains the distributions from mixtures on two of the resolved topologies. Polytopes (8), (9), and (10) correspond to mixtures on one, two, or three resolved topologies.

Polytopes (6) and (7) are of special interest, as they represent mixtures which are non-identifiable for phylogenetic reconstruction. In Observations 15 and 16 we are able to precisely delineate the mixtures which are non-identifiable across all three topologies rather than simply constructing examples as has been done previously [6, 13]. The drawback is that the mixtures found here may be on as many as eight sets of branch lengths (recall Theorem 1) rather than just two, and that we are mixing trees with extreme branch lengths.

There is one more polytope which we will investigate, which is that cut out by inequalities known to be satisfied for probability distributions of mixtures on trees. We will call this polytope  $L$ . Specifically,  $L$  is the polytope cut out by  $0 \leq q_A \leq 1$  for any  $A$ , and the Fourier transform of the inequalities  $0 \leq p_A \leq p_\emptyset$  for any  $A$  and the equality  $\sum_A p_A = 1$ . Note that the equality is equivalent to  $q_\emptyset = 1$ . The inequality in probability space is equivalent to  $\langle e_A, \underline{p} \rangle \geq 0$  which is equivalent to

$$\langle He_A, \underline{q} \rangle \geq 0. \tag{11}$$

The following observation notes further redundancies.

**Observation 14.**  $\langle He_A, \underline{q} \rangle \geq 0$  and  $q_A \geq 0$  for every split  $A$  implies  $q_\emptyset \geq q_A$  for every split  $A$ . These same hypotheses also imply that the corresponding

probability distribution on splits is “conservative,” i.e. that  $p_\emptyset \geq p_A$  for any  $A$ .

*Proof.* Assume there are  $n$  taxa. For the first assertion, let  $J$  be the  $n$  by  $n$  matrix with all entries one. Then  $J - H$  is a matrix with non-negative entries. Therefore  $\langle He_A, \underline{q} \rangle \geq 0$  for every split  $A$  implies that  $\langle H(J - H)e_A, \underline{q} \rangle \geq 0$  for every split  $A$ . But  $HJe_A = H\underline{1} = 2^{n-1}e_\emptyset$  and  $HH = 2^{n-1}I$ , giving the first assertion.

For the second assertion, note that  $He_\emptyset - He_A$  is a vector with positive entries, thus  $\langle He_\emptyset - He_A, \underline{q} \rangle$  is positive given the assumptions. Thus  $\langle e_\emptyset - e_A, \underline{p} \rangle \geq 0$ , which is equivalent to the second assertion.  $\square$

Because of these observations we note that  $L$  is the polytope in Fourier transform space cut out by  $q_A \geq 0$  and (11) for each  $A$ , as well as  $q_\emptyset = 1$ .

The following is a simple use of `polymake` to go from a  $V$ -representation to an  $H$ -representation.

**Observation 15.** *Co( $T_\star$ ) is defined by  $q_\emptyset = 1$ ,  $q_{123} \geq 0$  and the inequalities (11) and  $q_A \geq q_{123}$  for each  $A$ .*  $\square$

Another `polymake` calculation demonstrates

**Observation 16.** *The inclusion in (6) is an equality. In phylogenetic terms, the site pattern frequency vectors obtainable as a mixture of branch length sets on each of the three resolved quartet topologies are exactly those obtainable as a mixture of branch length sets on the four taxon star tree.*  $\square$

It is now easy to say what trees sit inside the star tree polytope  $Co(T_\star)$ .

**Proposition 17.** *The resolved quartet trees whose site pattern frequency vectors are obtainable as mixtures of branch length sets on the four taxon star tree are exactly those such that the internal branch length is shorter than the sum of the branch lengths for any two non-adjacent edges.*

This proposition may come as a surprise for phylogenetics researchers: even though a given data set may not have any evidence for a particular split, the data can appear to come from a tree with an internal edge which is longer than any of the pendant edges. Said another way, in order for the vector of expected site-pattern frequencies for a quartet tree to be identifiable, it is necessary that the internal edge be longer than the sum of the branch lengths for a pair of non-adjacent pendant edges.

*Proof.* Let  $\underline{q}$  denote the Fourier transform of the site pattern frequency vector for the tree in question, which we assume without loss of generality to have topology 12|34. This  $\underline{q}$  can be expressed as a mixture of branch lengths on the star tree exactly when it satisfies the conditions in Observation 15. Because  $\underline{q}$  is the Fourier transform of a vector in the probability simplex, by the above  $q_\emptyset = 1$ ,  $q_{123} \geq 0$ , and the inequality (11) is thus satisfied for any  $A$ . Now for each  $A \subset \{1, 2, 3\}$  we investigate the consequences of the inequality  $q_A \geq q_{123}$ . For  $A = \{1\}$ , the inequality becomes by (5)

$$\theta_1\theta_5\theta_4 \geq \theta_1\theta_2\theta_3\theta_4 \Leftrightarrow \theta_5 \geq \theta_2\theta_3.$$

Repeating the process for  $A = \{2\}, \{1, 3\}, \{2, 3\}$  and simplifying gives

$$\theta_5 \geq \max\{\theta_1\theta_3, \theta_1\theta_4, \theta_2\theta_3, \theta_2\theta_4\}.$$

The cases  $A = \{1, 2\}, \{3\}$  give  $1 \geq \theta_3\theta_4$  and  $1 \geq \theta_1\theta_2$ , which are trivially satisfied, as is the case of  $A = \{1, 2, 3\}$ . Taking logarithms and dividing by  $-2$  gives the desired result.  $\square$

In the previous section we showed that the dimension of the set of pattern probabilities which can be realized as a mixture of branch lengths on the star tree is equal to the dimension of those pattern probabilities which can be realized as an arbitrary mixture of trees. This means that given a sample from any nowhere-zero probability distribution on arbitrary phylogenetic mixtures there is a non-zero probability of having the sample be realizable from the set of mixture distributions on the star tree. However, it does not give any quantitative information. Quantitative answers for this and related questions for the uniform distribution on site-pattern frequencies can be calculated by using `polymake` to calculate volumes. Results are reported in Table 1.

For example, assume we choose uniformly a random probability distribution on patterns obtained by a mixture of trees of a given topology. Then there is a probability of approximately 0.57 ( $\approx 0.173/0.302$ ) that it is non-identifiable, i.e. that it can be written as a mixture of branch lengths on another topology. More work on the relevant geometry is needed to determine if such mixtures pose problems in the parameter regimes usually found in phylogenetics.

polytope	relative volume
$Co(T_*)$	0.142
$Co(T_1) \cap Co(T_2)$	0.173
$Co(T_1)$	0.302
$Co(T_1, T_2)$	0.565
$Co(T_1, T_2, T_3)$	0.908
$L$	1

Table 1: Relative volumes of the polytopes described in the text.

## 2 Mixtures of two trees

### 2.1 Combinatorics

In this section we establish a new combinatorial property that allows pairs of binary phylogenetic trees to be reconstructed from their induced subtrees of size at most six (Theorem 18). The statistical significance of this result is described in Corollary 20 and the next section. We begin with some definitions.

Let  $B(X)$  denote the collection of binary phylogenetic  $X$ -trees (up to isomorphism) and let  $B(X, k)$  denote the subsets of  $B(X)$  of size at most  $k$ . For  $T \in B(X)$  and  $Y \subseteq X$ , let  $T|_Y$  denote the induced binary phylogenetic  $Y$ -tree obtained from  $T$  by restricting the leaf set to  $Y$ . For  $\mathcal{P} = \{T_1, \dots, T_j\} \in B(X, k)$  let  $\mathcal{P}|_Y := \{T_1|_Y, \dots, T_j|_Y\} \in B(Y, k)$ . We will often stray from standard set theoretical notation when writing restrictions, for example  $T|_{\{a,b,c,d\}}$  will be written  $T|_{abcd}$ .

We say that a collection  $M$  of subsets of  $X$  *disentangles*  $B(X, k)$  if one can reconstruct any  $\mathcal{P}$  from the corresponding collection  $\{\mathcal{P}|_Y : Y \in M\}$ . This is

equivalent to the condition that for any pair  $\mathcal{P}, \mathcal{P}' \in B(X, k)$  we have

$$\mathcal{P} = \mathcal{P}' \Leftrightarrow \mathcal{P}|_Y = \mathcal{P}'|_Y \text{ for all } Y \in M.$$

If in addition, there is a polynomial time (in  $|X|$ ) algorithm that reconstructs  $\mathcal{P}$  from the set  $\{\mathcal{P}|_Y : Y \in M\}$  we say that  $M$  *efficiently disentangles*  $B(X, k)$ .

For example, it is well known that when  $k = 1$  the collection  $M$  of subsets of  $X$  of size four efficiently disentangles  $B(X, 1)(= B(X))$ ; indeed we may further restrict  $M$  to just those subsets of size four that contain a particular element, say  $x$ , of  $X$  (see, e.g., Theorem 6.8.8 of [10]). However, the subsets of  $X$  of size four do not suffice to disentangle  $B(X, 2)$ ; moreover, neither do the subsets of  $X$  of size at most five. To establish this last claim, let  $X = \{1, 2, \dots, 6\}$ , and consider two pairs of trees shown in Figure 2.1. Then  $\{T_1|_Y, T_2|_Y\} = \{T'_1|_Y, T'_2|_Y\}$  for all subsets  $Y$  of size at most five, yet  $\{T_1, T_2\} \neq \{T'_1, T'_2\}$ . However, allowing subsets of  $X$  of size at most six allows for the following positive result.

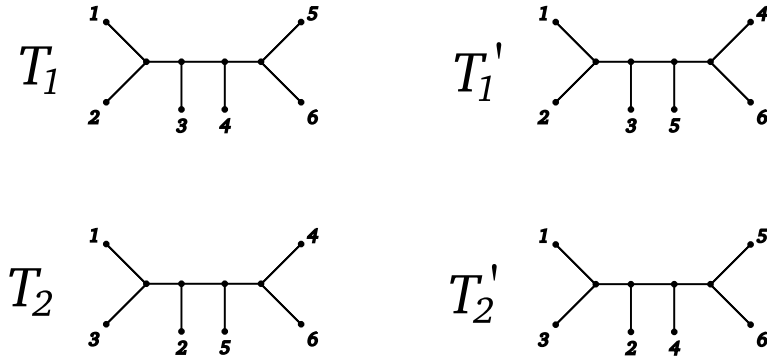


Figure 1: Two pairs of trees which have the same combined set of splits.

**Theorem 18.**  $B(X, 2)$  can be efficiently disentangled by the subsets of  $X$  of size at most six.

To establish this result we require the following lemma.

**Lemma 19.** Let  $T$  be a binary phylogenetic tree on a set  $Y$  of seven leaves, and suppose that  $S = \{a, b, c\}$  is a subset of  $Y$  of size three. Let  $x, y$  be any two distinct elements of  $Y - S$ . Then the quartet tree  $T|_{S \cup \{x\}}$  is determined by the collection of quartet trees  $T|_q$  as  $q$  ranges across the following four values:

(i)  $\{a, b, x, y\}, \{a, c, x, y\}, \{b, c, x, y\}$ , and

(ii)  $\{a, b, c, y\}$ .

*Proof.* Consider  $T|_{abcy}$ . Without loss of generality we may suppose that  $T|_{abcy} = ab|cy$ . If  $T|_{abxy} = ab|xy$  then  $T|_{S \cup \{x\}} = ab|cx$ . On the other hand, if  $T|_{abxy} = ax|by$  (or  $ay|bx$ ) then  $T|_{S \cup \{x\}} = ax|bc$  (or  $ac|bx$ , respectively).  $\square$

*Proof of Theorem 18.* Consider the collection  $Q$  of quartets of  $X$  that contain a given element  $x \in X$ . The quartets in  $Q$  are of two types: let  $Q_1$  denote the quartets  $q$  in  $Q$  for which  $T_1|_q = T_2|_q$  (i.e.  $\mathcal{P}|_q$  consists of just one tree) and let  $Q_2 = Q - Q_1$ . Set  $\mathcal{Q}_1 := \{T_1|_q (= T_2|_q) : q \in Q_1\}$  and set

$$\mathcal{Q}_2 := \{T_1|_q : q \in Q_2\} \cup \{T_2|_q : q \in Q_2\}.$$

From  $\mathcal{Q}_2$  we construct a graph  $G(\mathcal{Q}_2)$  that has vertex set  $\mathcal{Q}_2$  and that has an edge between two quartet trees, say  $ij|kl$  and  $i'j'|k'l'$ , precisely if one of the trees in  $\mathcal{P}$  displays both of these quartet trees. Note that  $G(\mathcal{Q}_2)$  is the disjoint

union of two cliques. Moreover, for any two quartets  $q, q' \in Q_2$ , each of the two trees in  $Q_2$  that correspond to  $q$  is adjacent (in  $G(Q_2)$ ) to precisely one of the two trees in  $Q_2$  that correspond to  $q'$ , and the resulting two edges form a matching for these four vertices.

Now, provided  $q \cup q'$  has cardinality at most six we can determine this matching since we can, by hypothesis, construct  $\mathcal{P}|_{q \cup q'}$  which must consist of two trees, and this pair of trees tells us how to match the two resolutions provided by  $\mathcal{P}$  for  $q$  (viz.  $\{T_1|_q, T_2|_q\}$ ) with the two resolutions of  $q'$  (viz.  $\{T_1|_{q'}, T_2|_{q'}\}$ ). In particular we can determine the two edges of  $G(Q_2)$  that connect these four vertices of  $G(Q_2)$ .

We claim that we can also determine (in polynomial time using just  $\mathcal{P}|_Y$  for choices of  $Y$  of size at most *six*) the matching between these four vertices of  $G(Q_2)$  in the remaining case where  $q \cup q'$  has cardinality seven.

Accepting for moment this claim, this allows us to reconstruct all the edges of  $G(Q_2)$  and in particular the two disjoint cliques of  $G(Q_2)$ , which bipartition  $Q_2$ . Taking the union of each clique with  $Q_1$  provides the pair of subsets  $\{\{T_1|_q : q \in Q\}, \{T_2|_q : q \in Q\}\}$  from which  $\{T_1, T_2\}$  can be recovered. Furthermore all of this can be achieved in polynomial time.

Thus it remains to establish the claim. Take two quartets  $q = \{a, b, c, x\}$  and  $q' = \{a', b', c', x\}$  from  $Q_2$  where we are assuming (since  $|q \cup q'| = 7$ ) that

$$\{a, b, c\} \cap \{a', b', c'\} = \emptyset.$$

We will now invoke Lemma 19 with  $S = \{a, b, c\}$  and  $Y = q \cup q'$ . Assume all of the four quartets in Lemma 19 are in  $Q_1$ ; by the conclusion of the lemma the quartet tree  $T|_{abcx}$  is uniquely determined. Thus  $\{a, b, c, x\} \in Q_1$ , which contradicts our assumption. Therefore at least one of the four quartets of type (i) or (ii) in Lemma 19 is in  $Q_2$ .

Suppose there exists a quartet  $q^*$  of type (i) in Lemma 19. Then  $q \cup q^*$  and  $q' \cup q^*$  both have cardinality at most six (for the latter, note that  $y$  in Lemma 19 must be one of the elements  $a', b', c'$  as  $y \in Y - q$ ) and so we can determine the matching. Similarly, since  $\{a', b', c', x\} \in Q_2$  we can invoke Lemma 19 with  $S = \{a', b', c'\}$  and the pair  $x, y'$  where  $y'$  is an element of  $Y - S$  different from  $x$ . By similar logic, at least one of the quartets satisfying condition (i) or (ii) in Lemma 19 must also be in  $Q_2$  for this choice of  $S$ . Once again if we can find a quartet satisfying condition (i) of Lemma 19 we can determine the matching. A remaining possibility is that in both cases (i.e. for  $S = \{a, b, c\}$  and  $S = \{a', b', c'\}$ ) we can only find a quartet in each case that satisfies condition (ii) of Lemma 19. Call these two quartets  $q_1 = \{a, b, c, y\}$  and  $q'_1 = \{a', b', c', y'\}$ , respectively. Then the three sets  $q \cup q_1$ ,  $q' \cup q'_1$  and  $q_1 \cup q'_1$  each have cardinality at most 6 (for the last case, note that  $y'$  is one of  $a, b, c$  and  $y$  is an element of  $a', b', c'$ ) and so we can determine the matching for these three pairs. This allows construction of  $T_i|_{q \cup q' \cup q_1 \cup q'_1}$  for  $i = 1, 2$  from the corresponding quartet trees; the matching for the four vertices of  $G(Q_2)$  corresponding to  $q \cup q'$  are then available by restriction. This completes the proof.  $\square$

An immediate consequence of Theorem 18 is the following.

**Corollary 20.** *Suppose a model has the property that from an arbitrary mixture of processes on two trees with the same leaf set of size six we can reconstruct the topology of the two trees. Then the same property applies for mixtures of two trees on any leaf set  $X$  (of any size greater than 6), and by an algorithm that is polynomial in  $|X|$ .*

**Remark** Peter Humphries has extended Theorem 18 to obtain analogous results for  $B(X, k)$  for  $k > 2$  (manuscript in preparation.)

## 2.2 Models

### Clocklike mixtures

Suppose one has a phylogenetic mixture on two trees  $T_1$  and  $T_2$ . In this section we are interested in whether one can reconstruct the pair  $\{T_1, T_2\}$  (or some information about this pair) from sufficiently long sequences. In the case where for each tree there is a stationary reversible Markov process (possibly also with rate variation across sites), and the (positive, finite) branch lengths of  $T$  satisfy a molecular clock some positive results are possible.

**Observation 21.** *The union of the splits in two trees  $T_1$  and  $T_2$  on the same taxon set can be recovered from a phylogenetic mixture on the two trees under a molecular clock.*

To see this we simply consider the function  $p : X \times X \rightarrow [0, 1]$  defined by setting  $p(x, y)$  to be the probability that species  $x$  and  $y$  are assigned different states by the mixture distribution (i.e.  $p(x, y)$  is the expected normalized

Hamming distance between the sequences). Then  $p = d_1 + d_2$  where (by the molecular clock assumption)  $d_1$  and  $d_2$  are monotone transformations of tree metrics realized by  $T_1$  and  $T_2$  respectively. By split decomposition theory ([2]) it follows that  $\Sigma(T_1) \cup \Sigma(T_2)$  can be recovered from  $p$ .

Note that  $\Sigma(T_1) \cup \Sigma(T_2)$  does not determine the set  $\{T_1, T_2\}$  as the two pairs of trees in Figure 2.1 shows. However this example is somewhat special:

**Lemma 22.** *Suppose  $\{T_1, T_2\}$  and  $\{T'_1, T'_2\}$  are two pairs of binary phylogenetic trees on the same set  $X$  of six leaves, and that*

$$\Sigma(T_1) \cup \Sigma(T_2) = \Sigma(T'_1) \cup \Sigma(T'_2).$$

*Then either  $\{T_1, T_2\} = \{T'_1, T'_2\}$  or the two pairs of trees are as shown in Figure 2.1 (up to symmetries).*

*Proof.* The proof is simply a case-by-case check of split compatibility graphs. A split compatibility graph is a graph where each split is represented by a vertex and an edge connects two splits which are compatible. In this case there are three nontrivial splits for each tree topology; three splits being realizable on a tree is equivalent to those three splits forming a clique in the split compatibility graph. Thus the lemma is equivalent to saying that up to symmetries there is only one subset of the vertices of the split compatibility graph for six taxa which can be expressed as two three-cliques in two different ways.

There are two unlabeled topologies on binary trees of six leaves: the caterpillar (with symmetry group of size eight) and the symmetric tree (with symmetry

group of size 48). First we divide the problem into the case of two caterpillar topologies, then the case of one caterpillar and one symmetric topology, finally two symmetric topologies. We label the two types of splits as follows: we call a split with three taxa on either side (such as 123|456) “type  $x$ ”, and a split with two taxa on one side and four on the other (such as 12|3456) “type  $y$ .”

Assume  $\{T_1, T_2\} \neq \{T'_1, T'_2\}$ . In the case of two caterpillar topologies it can be seen by eliminating cases that  $T_1$  and  $T_2$  cannot share a split of type  $y$ . Therefore the four type  $y$  splits of  $T_1$  and  $T_2$  must form a square of distinct vertices in the split compatibility graph. Further elimination shows that the two trees in Figure 2.1 are the only ones possible up to symmetries.

The cases involving a symmetric tree are even easier, as the choice of two splits in a symmetric tree determines the third. In the case of one caterpillar and one symmetric topology, this implies that there can be at most four type  $y$  splits in  $T_1$  and  $T_2$ . Checking cases quickly eliminates all possibilities. Similar reasoning deals with the two symmetric topology case, proving the lemma.

□

**Theorem 23.** *Suppose that for a reversible stationary model (possibly with rate variation across sites) there is a method that is able to distinguish a mixture of the two trees  $T_1$  and  $T_2$  from a mixture of trees  $T'_1$  and  $T'_2$  (see Figure 2.1) under branch lengths that satisfy a molecular clock on each tree. Then from any mixture on two binary trees on a leaf set  $X$  with both sets of branch lengths subject to a clock, one can recover the two trees by an algorithm that runs in polynomial time in  $|X|$ .*

*Proof.* Combine Theorem 18, Observation 21, and Lemma 22. □

### Non-clocklike mixtures

In [6] it was shown that under two-state symmetric (CFN) model one can have a mixture of two processes on one tree giving the same probability distribution on site patterns as a single process on a different tree. This requires that the two sets of branch lengths being mixed to be quite different and carefully adjusted. For example, we have:

**Corollary 24.** *If a mixture of two sets of branch lengths on a tree  $R$  has the same probability distribution on site patterns as a tree of a different topology  $S$ , then the two sets of branch lengths cannot be clock-like (even for different rootings of the tree), nor can one branch length set be a scalar multiple of the other.*

*Proof.* There must be a taxon set  $abcd$  such that  $R|_{abcd} = ab|cd$  and  $S|_{abcd} = ac|bd$ . Using the notation of [6], (also explained in Section 2.3) clocklike mixtures must have a pair of adjacent taxa (say  $a$  and  $b$ ) such that  $k_a = k_b$ . For one set of branch lengths to be a nontrivial scalar multiple of another, all of the pendant  $k_i$ 's must be either less than or greater than one. Either of these cases contradicts Proposition 7 of [6]. □

However, one could ask if a more complex mixture of branch lengths on one tree could mimic an unmixed process on a different tree. Again a molecular clock rules this out, and for branch lengths that scale proportionate (as in a rates-

across-sites distributions) we now show that identifiability of the underlying tree still holds.

**Theorem 25.** *Consider two binary phylogenetic trees  $T$  and  $T'$  on the same leaf set  $X$  of size  $n$  generating data under the CFN model. For  $T$  suppose we have a mixture of such processes that can be described by a set of branch lengths and a distribution  $\mathcal{D}$  of rates across sites which generates the same distribution on site patterns as that produced by an (unmixed) set of branch lengths on  $T'$ . Then  $T = T'$  and  $\mathcal{D}$  is the degenerate distribution that assigns all sites the same rate.*

*Proof.* It suffices to prove the result for  $n = 4$  and  $X = \{1, 2, 3, 4\}$ , with  $T$  the tree 12|34, and  $T'$  the tree 13|24. We denote the edge of  $T$  (resp.  $T'$ ) that is incident with leaf  $i$  by  $e_i$  (resp.  $e'_i$ ) and the interior edge of  $T$  (resp.  $T'$ ) by  $e_0$  (resp.  $e'_0$ ). Let  $\theta'_i := 1 - 2p(e'_i)$  and let  $\lambda_i$  denote the branch length of edge  $e_i$  so that the probability of a change along  $e_i$  is  $\frac{1}{2}(1 - f(2\lambda_i))$  where  $f(x) = \mathbb{E}_{\mathcal{D}}[\exp(\mu x)]$  is the moment generating function for the distribution of the rate parameter  $\mu$  in  $\mathcal{D}$ .

Then we have (see, e.g., Lemma 8.6.4 and Theorem 8.8.1 of [10]):

$$\theta'_1\theta'_2 = f(-2\lambda_1 - 2\lambda_2 - 2\lambda_0) \text{ and } \theta'_3\theta'_4 = f(-2\lambda_3 - 2\lambda_4 - 2\lambda_0),$$

and thus

$$f(-2\lambda_1 - 2\lambda_2 - 2\lambda_0) \cdot f(-2\lambda_3 - 2\lambda_4 - 2\lambda_0) = \theta'_1\theta'_2\theta'_3\theta'_4.$$

Also,

$$\theta'_1\theta'_3\theta'_2\theta'_4 = f(-2\lambda_1 - 2\lambda_2 - 2\lambda_3 - 2\lambda_4).$$

Combining these last two equations and setting  $r := -2\lambda_1 - 2\lambda_2, s := -2\lambda_3 - 2\lambda_4$ ;

$$f(r + s) = f(r - 2\lambda_0)f(s - 2\lambda_0) \leq f(r)f(s), \quad (12)$$

with equality precisely if  $\lambda_0 = 0$ . However,  $\exp(\mu x)$  is an increasing function of  $\mu$  for positive  $x$ . It follows that the random variables  $\exp(\mu r)$  and  $\exp(\mu s)$  are positively correlated, i.e.

$$f(r + s) \geq f(r)f(s)$$

with equality precisely if  $\mathcal{D}$  is a degenerate distribution. Consequently, (12) is an equality; thus  $\mathcal{D}$  is a degenerate distribution and  $T' = T$ .  $\square$

**Remark** Theorem 25 extends to provide an analogous result for the uniform distribution random cluster model on any even number  $q = 2r$  of states, since such a model induces the random cluster model on two states by partitioning the  $2r$  states into two sets, each of size  $r$ .

### 2.3 Mixed branch repulsion: larger trees

In this section we find results analogous to those in [6] for trees larger than quartet trees. The main result is that mixtures of two sets of branch lengths on a tree can only mimic a tree which is topologically one nearest neighbor interchange away from the original tree.

Let  $\ell(T)$  denote the set of leaves of  $T$ . We will write  $R \rightsquigarrow S$  to mean that there exists a mixture of two sets of branch lengths on  $R$  which produces exactly the same site pattern frequencies as some branch length set on a tree of topology  $S$  under the CFN model. Of course, if  $R \rightsquigarrow S$  then  $\ell(R) = \ell(S)$ .

**Theorem 26.** *Assume  $R$  and  $S$  are two topologically distinct trees on at least four leaves such that  $R \rightsquigarrow S$ . Then  $R$  and  $S$  differ topologically by one nearest neighbor interchange (NNI). Furthermore, assume the NNI partitions  $\ell(R)$  into the sets  $X_1, \dots, X_4$ . Then  $R|_{X_i} = S|_{X_i}$  for any  $i$  (equality as rooted trees with branch lengths).*

For this proof we will draw notation and several ideas from the proof of the main result of [6]. For a four taxon tree with taxon labels 1 through 4 we will label the the pendant edges with the corresponding numbers. We will write the quartet tree with the  $ab|cd$  split as simply  $ab|cd$ . Given two sets of branch lengths on a given tree we use  $k_i$  to denote the ratio of the fidelities (see Section 1.2) of the two branch lengths for the edge  $i$ . We will constantly use the simple fact that if the edge of an induced subtree consists of a sequence of edges then the induced  $k_i$  for that edge consists of the product of the  $k_i$ 's for the sequence of the edges.

**Lemma 27.** *The splits  $ab|cd$  and  $ac|bd$  are invariant under the Klein four group*

$$K_4 = \{1, (ab)(cd), (ac)(bd), (ad)(bc)\}.$$

□

The following lemma is easy to check.

**Lemma 28.** *Given numbers  $k_a, k_b, k_c$ , there exists  $\sigma \in K_4$  such that*

$$k_{\sigma(a)} \geq k_{\sigma(b)} \text{ and } k_{\sigma(a)} \geq k_{\sigma(c)}.$$

□

The following lemma is a rephrasing of Proposition 3 of [6]:

**Lemma 29.** *If  $ab|cd \rightsquigarrow ab|cd$  then the following two statements must be satisfied:*

- $k_a = k_b$  or  $k_c = k_d$
- $k_a = k_b^{-1}$  or  $k_c = k_d^{-1}$ .

□

**Lemma 30.** *If  $ab|cd \rightsquigarrow ac|bd$  then*

- *There is some element  $\sigma \in K_4$  such that  $k_{\sigma(a)} > k_{\sigma(c)} > k_{\sigma(d)} > k_{\sigma(b)}$*
- *none of  $k_a, \dots, k_d$  are equal to one*
- *exactly either one or three of  $k_a, \dots, k_d$  are greater than one*
- $k_a \neq k_b^{-1}$  and  $k_c \neq k_d^{-1}$ .

*Proof.* Each item in the list is from Proposition 7 of [6] with the exception of the last one. By Lemma 27 we can relabel such that  $k_a > k_c > k_d > k_b$ . Let

$f(x) = \frac{x^2-1}{x}$ . Note that  $f(x^{-1}) = -f(x)$ ,  $f(x)$  is positive for  $x \geq 1$  and strictly increasing for  $x > 0$ . By equation (12) of [6],

$$f(k_a)f(k_d) + f(k_b)f(k_c) > 0.$$

Assume first that  $k_a > k_c > k_d > 1 > k_b$ . Then, by the above properties of  $f$ ,

$$f(k_a)f(k_c) + f(k_b)f(k_c) > 0$$

$$f(k_a) + f(k_b) > 0$$

$$f(k_a) > f(k_b^{-1})$$

which implies  $k_a \neq k_b^{-1}$ . The case where  $k_a > 1 > k_c > k_d > k_b$  is similar, as is the proof that  $k_c \neq k_d^{-1}$ .  $\square$

The proof of Theorem 26 rests on the following easy observation.

**Lemma 31.** *If  $R \rightsquigarrow S$  then  $R|_F \rightsquigarrow S|_F$  for any  $F \subset \ell(R)$ .*  $\square$

We will use this lemma by restricting taxon sets of the larger tree to sets of size five, then analyzing for which ordered pairs  $(R, S)$  of five leaf subtrees it holds that  $R \rightsquigarrow S$ . There are 225 ordered pairs of five leaf trees, however in the following lemma we show that symmetry considerations reduce the relevant number of interest to four. For ease of notation, we will write the five leaf subtree  $W_{abcde}$  as shown in Figure 2.

**Lemma 32.** *Given trees on five leaves  $R$  and  $S$ , the question of whether  $R \rightsquigarrow S$*

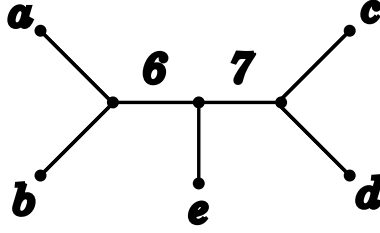


Figure 2: Definition of  $W_{abcde}$ .

or not is equivalent to the question of if one of the following is true:

$$W_{12345} \rightsquigarrow W_{12345} \tag{13}$$

$$W_{12345} \rightsquigarrow W_{13245} \tag{14}$$

$$W_{12345} \rightsquigarrow W_{12354} \tag{15}$$

$$W_{12345} \rightsquigarrow W_{13254} \tag{16}$$

*Proof.* Clearly it can be assumed that  $R$  is  $W_{12345}$  by renumbering. Note that the symmetries of a five leaf tree are generated by (12), (34), and (13)(24) on the tree  $W_{12345}$ . A combination of these symmetries applied to  $R$  and renumbering means that these symmetries can then be applied to  $S$  while still assuming that  $R$  is  $W_{12345}$ . Using these symmetries  $S$  can be assumed to be either  $W_{abcd4}$  or  $W_{abcd5}$ . There are six such trees; a further application of the symmetries shows that the cases of  $S = W_{13254}$  and  $S = W_{23154}$  are equivalent, as are  $S = W_{13245}$  and  $S = W_{14235}$ .  $\square$

**Lemma 33.** *Mixture (14) is impossible, i.e.  $W_{12345} \not\rightsquigarrow W_{13245}$ .*

*Proof.* Assume the contrary, and that  $k_i$ 's are labeled as in Figure 2. By (clear extensions of) Lemmas 27 and 28 we can assume that  $k_1 \geq k_2$  and  $k_1 \geq k_3$  on these trees. By restricting to the taxon set to 1234, and noting that by Lemma 31  $12|34 \mapsto 13|24$ , we have  $k_1 > k_3 > k_4 > k_2$  and that  $k_3$  and  $k_4$  are either both greater than one or both less than one by Lemma 30. By restricting to 1235, it is clear that  $k_5 \neq 1$ . Assume  $k_5 < 1$ . Restricting the taxon set to 2345 means that  $25|34 \mapsto 24|35$ ; by testing elements of  $K_4$  in Lemma 30 and using the fact that  $k_3$  and  $k_4$  are either both greater than one or both less than one and that  $k_5 < 1$ , one must have  $k_2k_6 > k_4 > k_3 > k_5$ . This contradicts the above statement that  $k_3 > k_4$ . The case where  $k_5 > 1$  follows similarly by restricting to 1345.  $\square$

**Lemma 34.** *Mixture (16) is impossible, i.e.  $W_{12345} \not\rightarrow W_{13254}$ .*

*Proof.* Assume the contrary. First restrict to the taxon set 1345. For this taxon set  $15|34 \mapsto 13|45$ , showing by Lemma 30 that  $k_3 \neq k_4$ ,  $k_3 \neq k_4^{-1}$ , and  $k_5 \neq 1$ . Second, restrict to taxon set 2345. For this taxon set the induced mixture is  $25|34 \mapsto 25|34$ , therefore we apply Lemma 29. Because  $k_3 \neq k_4$  and  $k_3 \neq k_4^{-1}$ , it must be true that  $k_2k_6 = k_5$  and  $k_2k_6 = k_5^{-1}$ . This contradicts the fact that  $k_5 \neq 1$ .  $\square$

Therefore we are left with mixtures (13) and (15), implying the following corollary.

**Corollary 35.** *Assume  $R \mapsto S$  for two five-leaf trees  $R$  and  $S$ . Then  $R$  and  $S$  share a nontrivial split.*  $\square$

We now present two more lemmas which will be used in the proof of Theorem 26. Given rooted trees  $R$  and  $S$  let  $R-S$  denote the unrooted tree obtained by joining the roots of  $R$  and  $S$  together with an edge.

**Lemma 36.** *Assume  $R_1 - R_2 \rightsquigarrow S_1 - S_2$ ,  $\ell(R_1) = \ell(S_1)$ , and all of the  $k$ 's for the edges in  $R_1$  are one. Then  $R_1 = S_1$  (equality with branch lengths).*

*Proof.* Add a taxon  $e$  at the root of  $R_1$  (resp.  $S_1$ ) to obtain the unrooted tree  $R_U$  (resp.  $S_U$ ). We will show that the between-leaf distance matrices for  $R_U$  and  $S_U$  are the same, which implies that  $R_U = S_U$  and thus  $R_1 = S_1$ . Pick  $c$  and  $d$  distinct in  $\ell(R_2)$ . Pick an arbitrary  $a$  and  $b \in \ell(R_1)$  and restrict to the taxon set  $abcd$ . By Proposition 4 of [6], the pairwise distance between  $a$  and  $b$  in  $R_1$  and  $S_1$  (and thus in  $R_U$  and  $S_U$ ) will be the same. To show that distances from taxa  $a \in \ell(R_1)$  to the root taxon  $e$  are the same in  $R_U$  and  $S_U$ , repeat the same process but for any  $a$  choose  $b$  such that the MRCA of  $a$  and  $b$  in  $R_1$  is the root of  $R_1$ . Another application of Proposition 4 of [6] in this case proves the proposition. □

**Lemma 37.** *If  $R_1 - R_2 \rightsquigarrow S_1 - S_2$ ,  $\ell(R_1) = \ell(S_1)$  and  $\Sigma(R_2) \neq \Sigma(S_2)$  then  $R_1 = S_1$  (equality with branch lengths).*

*Proof.* For  $x, y \in \ell(R_2)$ , let  $C_y(x)$  be the set of edges in the path from  $x$  to the MRCA of  $x$  and  $y$ . Define

$$\varphi_y(x) = \prod_{e \in C_y(x)} k_e.$$

This takes the place (for induced subtrees) of a single  $k_e$ . The idea of the proof is to use the previous lemma by showing that  $k_e$  for any edge  $e$  in  $R_1$  is one. However, by induction it is enough to show that  $\varphi_y(x) = \varphi_x(y) = 1$  for any  $x, y \in \ell(R_2)$ .

Since  $\Sigma(R_2) \neq \Sigma(S_2)$  but  $\ell(R_2) = \ell(S_2)$  there exists a subset  $\{a, b, c\} \subset \ell(R_2)$  such that  $R_2$  restricted to the taxon set  $abc$  is the tree  $(ab)c$ , while  $S_2$  restricted to  $abc$  is  $(ac)b$ . Pick any  $x, y \in \ell(R_1)$ . First restrict to taxon set  $abcx$ , for which  $ab|cx \rightarrow ac|bx$ . By Lemma 30,  $\varphi_b(a) \neq \varphi_a(b)$  and  $\varphi_b(a) \neq [\varphi_a(b)]^{-1}$ . Now restrict to the taxon set  $abxy$ , for which  $ab|xy \rightarrow ab|xy$ . By Lemma 29,  $\varphi_y(x) = \varphi_x(y)$  and  $\varphi_y(x) = [\varphi_x(y)]^{-1}$ , implying that each  $\varphi$  is one. The lemma now follows.  $\square$

The final lemma allows for the combination of splits; it is a special case of Lemma 2 of [7]. An easy argument is presented here.

**Lemma 38.** *Let  $T$  be a phylogenetic tree. If  $A \cup \{x\}|B \in \Sigma(T|_{A \cup B \cup \{x\}})$  and  $A \cup \{y\}|B \in \Sigma(T|_{A \cup B \cup \{y\}})$  then  $A \cup \{x, y\}|B \in \Sigma(T|_{A \cup B \cup \{x, y\}})$ .*

*Proof.* First we note that if  $A|B \in \Sigma(T|_{A \cup B})$  then one of  $A|B \cup \{x\}$  or  $A \cup \{x\}|B$  is contained in  $\Sigma(T|_{A \cup B \cup \{x\}})$ , otherwise the restriction of  $T|_{A \cup B \cup \{x\}}$  to  $A \cup B$  cannot contain the split  $A|B$ .

Applying this fact to the two splits  $A \cup \{x\}|B$  and  $A \cup \{y\}|B$  implies either the conclusion of the lemma or that  $A \cup \{x\}|B \cup \{y\}$  and  $A \cup \{y\}|B \cup \{x\}$  are both in  $\Sigma(T|_{A \cup B \cup \{x, y\}})$ . This latter option is excluded by split compatibility.  $\square$

*Proof of Theorem 26.* Because  $R$  and  $S$  are topologically distinct yet have the

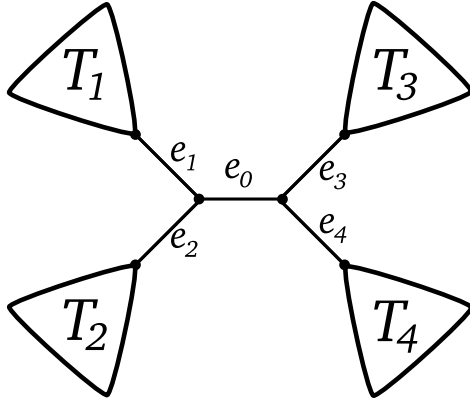


Figure 3: Notation used in the proof of Theorem 26.

same number of leaves, there must be at least one split in  $R$  which is not in  $S$ . Say this split is given by the edge  $e_0$ . The edge  $e_0$  must induce a nontrivial split, and therefore assign  $e_1, \dots, e_4$  and  $T_1, \dots, T_4$  such that  $R$  can be drawn as in Figure 3.

Pick any  $i \in \{1, \dots, 4\}$ . We claim that the split induced by edge  $e_i$  is in  $\Sigma(S)$ . If  $|\ell(T_i)| = 1$  then there is nothing to prove, so assume that  $|\ell(T_i)| \geq 2$ . Construct a five-leaf tree by choosing two leaves  $a, b$  from  $\ell(T_i)$  and also leaves  $c, d, e$ : one from each of the other three  $T_j$ . Because the split induced by  $e_0$  is not in  $S$  by hypothesis, it also cannot be in  $S|_{abcde}$ . An application of Corollary 35 now implies that the split induced by  $e_i$  must be in  $\Sigma(S|_{abcde})$ . This is true for each such choice of  $abcde$ : of these choices combined via Lemma 38 show that the split induced by the edge  $e_i$  is in  $\Sigma(S)$ .

Four applications of Lemma 37 now prove the theorem. □

The following proposition says that the sort of mixture described in The-

orem 26 is possible (assuming the main result of [6]). It is a simple general fact.

**Proposition 39.** *Let  $T_1, \dots, T_4$  be rooted trees and  $R$  and  $S$  two trees on the taxon set 1234. Let  $\tilde{R}$  and  $\tilde{S}$  be the trees obtained from  $R$  and  $S$  by attaching tree  $T_i$  to taxon  $i$ . Now if  $R \rightsquigarrow S$  then  $\tilde{R} \rightsquigarrow \tilde{S}$ .*

*Proof.* Let the vector  $\underline{y}$  represent the state vector for the terminal taxa on  $R$  and  $S$  and let  $\underline{x}_i$  represent the state vector for the tree  $T_i$ . Let  $p_\gamma^T(\underline{z})$  mean the probability of state vector  $\underline{z}$  on a tree with branch lengths  $\gamma$ ;  $\gamma$  will be omitted if understood. The statement  $R \rightsquigarrow S$  means exactly that there exist  $\gamma_1, \gamma_2, \gamma_3$  and  $\alpha$  such that

$$\alpha p_{\gamma_1}^R(\underline{y}) + (1 - \alpha) p_{\gamma_2}^R(\underline{y}) = p_{\gamma_3}^S(\underline{y})$$

for any state vector  $\underline{y}$ . We observe that

$$p^{\tilde{W}}(\underline{x}_1, \dots, \underline{x}_4) = \sum_{\underline{y}} p^W(\underline{y}) \prod_{i=1}^4 p^{T_i}(\underline{x}_i | \underline{y})$$

for  $W = R, S$ , which implies

$$\begin{aligned} \alpha p_{\tilde{\gamma}_1}^{\tilde{R}}(\underline{x}_1, \dots, \underline{x}_4) + (1 - \alpha) p_{\tilde{\gamma}_2}^{\tilde{R}}(\underline{x}_1, \dots, \underline{x}_4) \\ = \sum_{\underline{y}} (\alpha p_{\gamma_1}^R(\underline{y}) + (1 - \alpha) p_{\gamma_2}^R(\underline{y})) \prod_{i=1}^4 p^{T_i}(\underline{x}_i | \underline{y}) \\ = p_{\tilde{\gamma}_3}^{\tilde{S}}(\underline{x}_1, \dots, \underline{x}_4) \end{aligned}$$

where the  $\tilde{\gamma}_j$  are simply the  $\gamma_j$  along with the branch lengths of the  $T_i$ .

□

For completeness we also record when a mixture of two sets of branch lengths can mimic a tree of the same topology under the CFN model.

**Proposition 40.** *If two sets of branch lengths on a tree mix to mimic a tree of the same topology under the binary symmetric model, then all branch lengths between the two sets must be the same with the possible exception of those for a quartet of adjacent edges sitting inside the tree.*

*Proof.* Assume a counter-example to Proposition 40: i.e. that there exists a tree  $R$  with two branch length sets which differ by more than a quartet of adjacent edges but which mix to mimic a tree of the same topology  $S$  under the binary symmetric model. Therefore, there exists a partitioning of  $R$  into subtrees  $A$ ,  $B$ , and  $C$  meeting at a node such that there is an edge in each of  $A$  and  $B$  which differs in terms of branch length. Note that if two branch length sets differ on a nontrivial rooted tree, then by induction one can find an induced rooted subtree of size two which differs in terms of branch length between the two branch length sets. Therefore there must be an induced rooted subtree of size two in each of  $A$  and  $B$  which differs in terms of branch length between the two branch length sets. Number the taxa thus chosen from  $A$  1 and 2, and the taxa chosen from  $B$  3 and 4. Label an arbitrary taxon from  $C$  with 5. Now consider the 5-taxon tree induced by restricting the taxon set to 1 through 5. Label the edges as in Figure 2, and assign  $k_i$ 's as before.

From the above we can assume (perhaps after renumbering) that  $k_1 \neq 1$  and  $k_3 \neq 1$ . By restricting  $R$  to the taxon set 1234 we have by Lemma 29 that  $k_1 = k_2^{-1}$  and  $k_3 = k_4$  (perhaps after renumbering.) Because  $k_3 \neq 1$ , clearly

$k_3 \neq k_4^{-1}$ . Thus using Lemma 29, by restricting to 1534 we have  $k_1 k_6 = k_5$  and by restricting to 2534 we have  $k_2 k_6 = k_5$ . Therefore  $k_1 = k_2 = 1$ , which is a contradiction.  $\square$

### 3 Conclusion

In conclusion, we have presented a number of new results which help to clarify when non-identifiable mixtures may pose a problem for reconstruction. However, the message isn't completely straightforward. The first section shows that the space of site pattern frequencies for mixtures of many trees contains a relatively large non-identifiable region. Furthermore, this non-identifiable region contains site-pattern frequencies for resolved trees with substantial internal branch lengths. Yet, these spaces were constructed using specific trees of extreme branch lengths, raising the question of whether corresponding results hold for more reasonable parameter regimes and "random" sets of trees which one might find from data. Also, some of the results hold only for the CFN model. Furthermore, we wonder if it is possible to find simple  $H$ -descriptions of the mixture polytope for larger star trees.

On the other hand, the second section shows generally that mixtures of just two trees may not pose so much of a problem. In particular, our results make progress towards showing that clocklike mixtures of pairs of branch lengths may be identifiable under further assumptions. We also show that pairs of trees under CFN rates-across-sites mixtures are identifiable. Finally, we show that

mixtures of pairs of branch lengths on a tree cannot “change” the topology too much.

In general, many interesting questions remain and we look forward to seeing further progress in this field.

### Acknowledgements

The authors are grateful to Sebastien Roch for helpful discussions on this work.

### References

- [1] E. S. Allman and J. A. Rhodes. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *J. Comput. Biol.*, 13(5):1101–1113, June 2006.
- [2] H. J. Bandelt and A. W. M. Dress. A canonical decomposition theory for metrics on a finite set. *Adv. Math.*, 92:47–105, 1992.
- [3] J. Felsenstein. *Inferring Phylogenies*. Sinauer Press, Sunderland, MA, 2004.
- [4] E. Gawrilow and M. Joswig. Geometric reasoning with polymake. [arXiv:math.CO/0507273](https://arxiv.org/abs/math/0507273), 2005.
- [5] B. Grünbaum. *Convex Polytopes*. Springer-Verlag, Berlin, 2003.
- [6] F. A. Matsen and M. Steel. Phylogenetic mixtures on a single tree can mimic a tree of another topology. [arXiv:0704.2260v1 \[q-bio.PE\]](https://arxiv.org/abs/0704.2260v1), 2007.
- [7] C. A. Meacham. Theoretical and computational considerations of the compatibility of qualitative taxonomic characters. In J. Felsenstein, editor, *Numerical taxonomy*, volume G1 of *NATO ASI Series*, pages 304–314. Springer-Verlag, Berlin, 1983.
- [8] H. Ochman, J. G. Lawrence, and E. A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, May 2000.
- [9] A. Rokas, B. L. Williams, N. King, and S. B. Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960):798–804, Oct 2003.

- [10] C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.
- [11] C. Simon, L. Nigro, J. Sullivan, K. Holsinger, A. Martin, A. Grapputo, A. Franke, and C. McIntosh. Large differences in substitutional pattern and evolutionary rate of 12S ribosomal RNA genes. *Mol Biol Evol*, 13(7):923–932, Sep 1996.
- [12] M. A. Steel, L. A. Szekely, and M. D. Hendy. Reconstructing trees when sequence sites evolve at variable rates. *J. Comput. Biol.*, 1(2):153–163, Summer 1994.
- [13] D. Štefankovič and E. Vigoda. Phylogeny of mixture models: Robustness of maximum likelihood and non-identifiable distributions. *J. Comput. Biol.*, 14(2):156–189, 2007.
- [14] G. M. Ziegler. *Lectures on Convex Polytopes*. Springer-Verlag, Berlin, 1994.