Enumeration of binary trees compatible with a perfect phylogeny

Julia A. Palacios[*]
Anand Bhaskar[†]
Filippo Disanto[‡]
Noah A. Rosenberg[§]

August 12, 2021

Abstract

Evolutionary models used for describing molecular sequence variation suppose that at a non-recombining genomic segment, sequences share ancestry that can be represented as a genealogy—a rooted, binary, timed tree, with tips corresponding to individual sequences. Under the infinitely-many-sites mutation model, mutations are randomly superimposed along the branches of the genealogy, so that every mutation occurs at a chromosomal site that has not previously mutated; if a mutation occurs at an interior branch, then all individuals descending from that branch carry the mutation. The implication is that observed patterns of molecular variation from this model impose combinatorial constraints on the hidden state space of genealogies. In particular, observed molecular variation can be represented in the form of a perfect phylogeny, a tree structure that fully encodes the mutational differences among sequences. For a sample of n sequences, a perfect phylogeny might not possess n distinct leaves, and hence might be compatible with many possible binary tree structures that could describe the evolutionary relationships among the n sequences. Here, we investigate enumerative properties of the set of binary ranked and unranked tree shapes that are compatible with a perfect phylogeny, and hence, the binary ranked and unranked tree shapes conditioned on an observed pattern of mutations under the infinitely-many-sites mutation model. We provide a recursive enumeration of these shapes. We consider both perfect phylogenies that can be represented as binary and those that are multifurcating. The results have implications for computational aspects of the statistical inference of evolutionary parameters that underlie sets of molecular sequences.

1 Introduction

Coalescent and mutation models are used in population genetics to estimate evolutionary parameters from samples of molecular sequences (Marjoram and Tavaré, 2006). The central idea is that observed molecular variation is the result of a process of mutation along the branches of the genealogy of the sample. This genealogy is a timed tree that represents the ancestral relationships of

---

[*]Department of Statistics and Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. Corresponding author: juliapr@stanford.edu

[†]Department of Genetics, Stanford University, Stanford University, Stanford, CA, USA.

[‡]Department of Mathematics, University of Pisa, Pisa, Italy.

[§]Department of Biology, Stanford University, Stanford, CA, USA.

the sample at a chromosomal segment. Consisting of a tree topology and its branch lengths, the genealogy is a nuisance parameter that is modeled as a realization of the coalescent process dictated by evolutionary parameters—which are in turn inferred by integrating over the space of genealogies. For large sample sizes, however, this integration is computationally challenging because the state space of tree topologies increases exponentially with the number of sampled sequences.

Recently, a coarser coalescent model known as the Tajima coalescent (Tajima, 1983; Sainudiin et al., 2015), coupled with the infinitely-many-sites mutation model (Kimura, 1969) has been introduced for population-genetic inference problems (Palacios et al., 2019). Whereas the standard coalescent model (Kingman, 1982) induces a probability measure on the space of ranked labeled tree topologies, the Tajima coalescent induces a probability measure on the space of ranked unlabeled tree topologies. Removing the labels of the tips from the tree topology, as in the Tajima coalescent, reduces the cardinality of the space of tree topologies substantially, shrinking computation time in inference problems.

Under infinitely-many-sites mutation, only a subset of tree topologies (labeled or unlabeled) are compatible with an observed data set, so that the computational complexity of inference varies among different data sets. Hence, Cappello et al. (2020a) used importance sampling to approximate cardinalities of the spaces of labeled and unlabeled ranked tree shapes conditioned on a data set of molecular sequences, demonstrating a striking reduction of the cardinality of the space of ranked unlabeled tree shapes versus the labeled counterpart when conditioning on observed data with a sparse number of mutations. Here, we extend beyond the approximate work of Cappello et al. (2020a) and obtain exact results. We provide a recursive algorithm for exact computation of the cardinality of the spaces of labeled and unlabeled ranked tree shapes compatible with a sequence data set. We provide a number of other enumerative results relevant for inference of tree topologies in phylogenetics and population genetics. Python code for enumeration is available at
`https://colab.research.google.com/drive/1cAx2xyn7OtmG-F-9nxJ3CHRc7e7AjuCj?usp=sharing`.

2 Preliminaries

2.1 Types of trees

The **coalescent** is a continuous-time Markov chain with values in the space $\mathcal{P}_n$ of partitions of $[n] = \{1, 2, \ldots, n\}$ (Kingman, 1982). The process starts with the trivial partition of n singletons, labeled $\{1\}, \{2\}, \ldots, \{n\}$, at time 0; at each transition, two blocks are chosen uniformly at random to merge into a single block. The process ends with a single block with label $\{1, 2, \ldots, n\}$. In the standard coalescent, the holding times are exponentially distributed with rate $\binom{k}{2}$ when there are k blocks. Transition probabilities for the coalescent can be factored into two independent components, a pure death process and a discrete jump chain. A full realization of the process can be represented by a timed rooted binary tree: a genealogy. The tips of the genealogy are labeled by $\{1, 2, \ldots, n\}$. Figure 1A shows a realization of the jump process, a ranked labeled tree shape.

A lumping of the standard coalescent process, called the **Tajima coalescent** (Sainudiin et al., 2015), consists in removing the labels of the tips of the genealogy. The pure death process of the lumped process is the same as the standard coalescent. The discrete jump chain can be described as a simple urn process (Janson and Kersting, 2011). Start with an urn of n balls labeled 0; at the ith transition, draw two balls and return one to the urn with label i. The process ends when there is a single ball with label $n-1$ in the urn. A full realization of the urn process can be represented

(A) Ranked and Labeled   (B) Ranked and Unlabeled   (C) Unranked and Unlabeled   (D) Unranked and Labeled
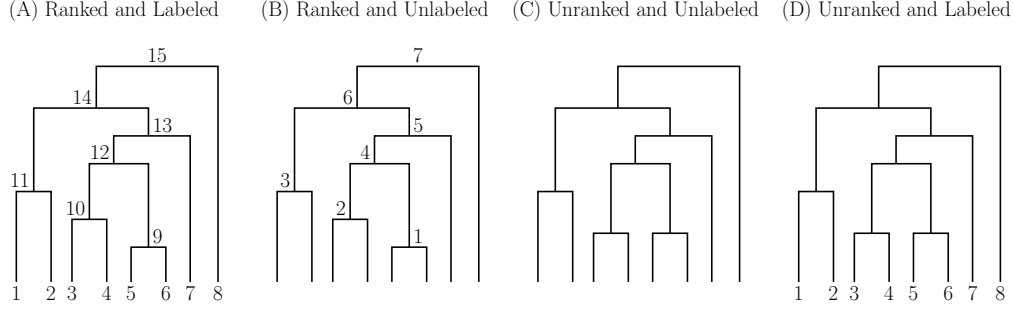
Figure 1: **Different types of trees.** (A) A ranked labeled tree shape. (B) A ranked unlabeled tree shape. (C) An unranked unlabeled tree shape. (D) An unranked labeled tree shape. The ranked unlabeled tree shape in (B) is obtained by discarding leaf labels from the ranked labeled tree shape in (A). The unranked labeled tree shape in (D) is obtained by discarding the sequence of internal node ranks in (A). The unranked unlabeled tree shape in (C) is obtained by discarding the sequence of internal node ranks in (B) or the leaf labels in (D).

as a ranked unlabeled tree shape with internal nodes labeled by the transition index.

A **ranked labeled tree shape** of size n, denoted by T_n^L, is a rooted binary labeled tree of n leaves with a total ordering for the internal nodes. Without loss of generality, we use label set $[n]$ to label the n leaves. The space of ranked labeled tree shapes with n leaves will be denoted by $\mathcal{T}_n^L$. Figure 1A shows an example of a ranked labeled tree shape with $n = 8$ leaves. Ranked labeled tree shapes are also known as labeled histories.

A **ranked unlabeled tree shape** of size n, denoted by T_n^R, is a rooted binary unlabeled tree of n leaves with a total ordering for the internal nodes. The space of ranked unlabeled tree shapes with n leaves will be denoted by $\mathcal{T}_n^R$. Figure 1B shows an example of a ranked unlabeled tree shape with $n = 8$ leaves. We will refer to a ranked unlabeled tree shape simply as a ranked tree shape; these ranked tree shapes are also known as unlabeled histories, or Tajima trees. Figure 2 shows all ranked unlabeled tree shapes with $3, 4, 5$, and 6 leaves.

An **unranked unlabeled tree shape** of size n, denoted by T_n, is a rooted binary unlabeled tree of n leaves with unlabeled internal nodes. The space of unranked (unlabeled) tree shapes with n leaves will be denoted by $\mathcal{T}_n$. Figure 1C shows an example of an unranked unlabeled tree shape with $n = 8$ leaves. These shapes are also called unlabeled topologies or Otter trees (Otter, 1948).

An **unranked labeled tree shape** of size n, denoted by T_n^X, is a rooted binary labeled tree of n leaves with unlabeled internal nodes. The space of unranked labeled tree shapes with n leaves will be denoted by $\mathcal{T}_n^X$. Figure 1D shows an example of an unranked labeled tree shape with $n = 8$ leaves. These tree shapes are also called labeled topologies.

2.2 Mutations on trees

Many generative models of neutral molecular evolution assume that a process of mutations is superimposed on the genealogy as a continuous-time Markov process. In the **infinitely-many-sites mutation model**, every mutation along the branches of the tree occurs at a chromosomal site that has not previously mutated (Kimura, 1969). Therefore, if a mutation occurs at an interior branch along the genealogy, all sequences descended from that branch carry the mutation. Because every site can mutate at most once, the sequence of mutated sites can be encoded as a binary
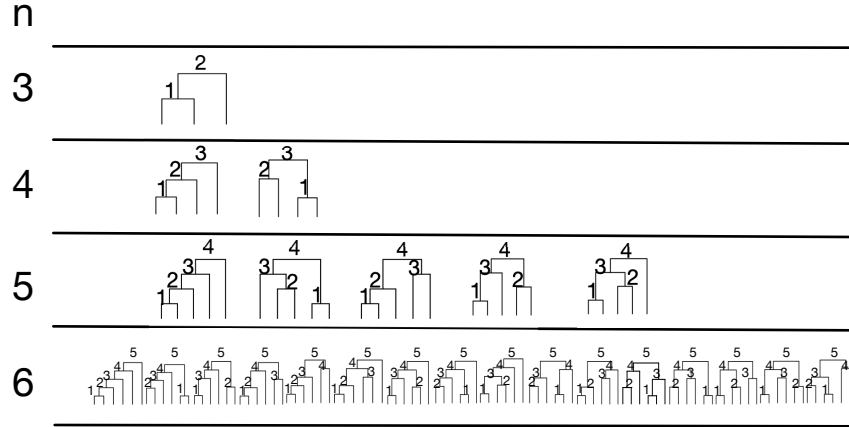
3

Figure 2: **An enumeration of all possible ranked tree shapes with 3, 4, 5, and 6 leaves.**

sequence, with 0 denoting the ancestral type and 1 denoting the mutant type at any site.

Figure 3A shows a realization of the Tajima coalescent together with a realization of mutations from the infinitely-many-sites mutation model with 5 individuals and 4 mutated sites. In what follows, we assume that we observe molecular data only as binary sequences at the tips of the tree.

2.3 Observed binary molecular sequence data as a perfect phylogeny

The **perfect phylogeny algorithm**, proposed by Gusfield (1991), generates a graphical representation of binary molecular sequence data that have been produced according to the infinitely-many-sites mutation model. Label individual sequences $1, 2, \ldots, n$, and label mutated or "segregating" sites $a, b, \ldots$. The original algorithm generates a tree structure known as a **perfect phylogeny**, with tips labeled $1, 2, \ldots, n$ and with edges labeled $a, b, \ldots$, that is in bijection with the observed "labeled data." An edge can have no labels, one label, or more than one label. Perfect phylogenies have been central to coalescent-based inference algorithms, in which maximum likelihood or Bayesian estimation of evolutionary parameters that have given rise to the particular distribution of mutations and clade sizes on the perfect phylogeny are sought by importance sampling or Markov chain Monte Carlo (Griffiths and Tavaré, 1994; Stephens and Donnelly, 2000; Palacios et al., 2019; Cappello et al., 2020b).

In this study, we assume that individual sequences are not uniquely labeled, but instead, are identified by their sequences of 0s and 1s, or **haplotypes**. Hence, the number of tips in our perfect phylogeny is the number of unique haplotypes, and the labels at the tips correspond to the observed frequencies of the haplotypes. For the genealogy in Figure 3A, Figure 3B shows the perfect phylogeny of the data observed at its tips.

The key assumption of the bijection between sequence data sets and perfect phylogenies is that if a site mutates once, then all descendants of the lineage on which the mutation occurred must also have the mutation—and no other individuals will have the mutation. That is, every unique mutation, or site, partitions the sample of haplotypes into two groups: those with the mutation and those without the mutation. Hence, we group sites that induce the same partition on the haplotypes, and we call each such group of sites a **mutation group**.
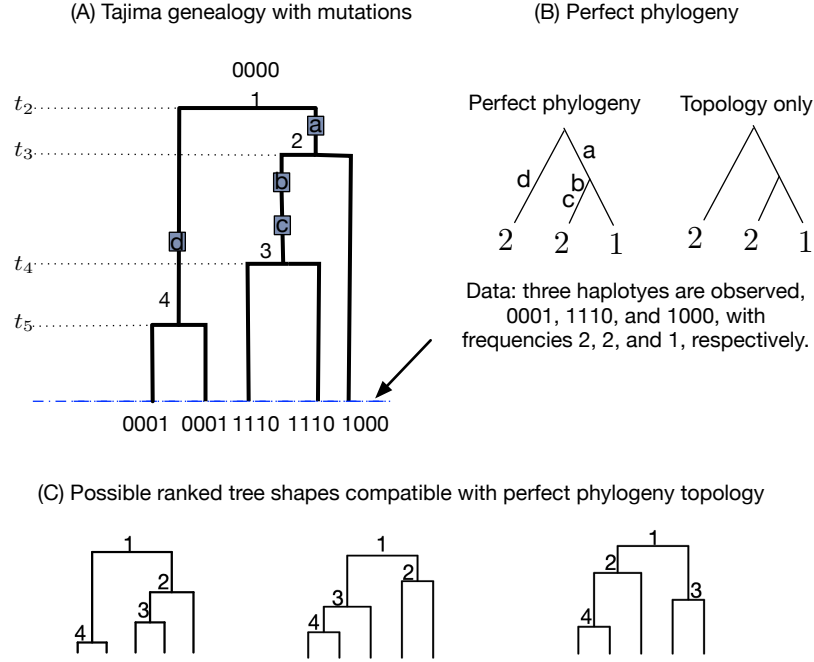
4

Figure 3: **Tajima coalescent and infinitely-many-sites generative model of binary molecular data.** **(A)** A Tajima genealogy of 5 individuals, with 4 superimposed mutations depicted as gray squares. The root is labeled by the ancestral type 0000, and the leaves are labeled by the genetic type at each of three mutated sites. The first two leaves from left to right are labeled 0001 because one mutation occurs in their path to the root. The third and fourth individuals have three mutations in their path to the root and are labeled 1110; the last individual is labeled 1000 because only one mutation occurs along its path to the root. The order and label of the mutations is unimportant; however, it is assumed that the same position, or site, in a sequence of 0s and 1s corresponds across individuals. For ease of exposition, we label the mutations a, b, c and d. The first site corresponds to mutation a, the second to b, the third to c, and the fourth to d. **(B)** Left, a perfect phylogeny representation of the observed data at the tips of (A). Data consist of 3 unique haplotypes 0001, 1110 and 1000, with frequencies 2, 2, and 1. The corresponding frequencies are the labels of tips of the perfect phylogeny. Right, perfect phylogeny topology obtained by removing the edge labels of the perfect phylogeny. **(C)** The only three ranked tree shapes compatible with the perfect phylogeny topology in (B).

In this study, we are not concerned with the mutation labels, and hence, we remove the edge labels of the perfect phylogeny (right side of Figure 3B), so that we consider only the topology of the perfect phylogeny. In dropping the edge labels, we treat a perfect phylogeny topology as a perfect phylogeny. Henceforth, a **perfect phylogeny** is a multifurcating rooted tree with k leaves, representing k distinct haplotypes, each labeled by a positive integer $(n_i)_{1 \leq i \leq k}$, with $\sum_{i=1}^{k} n_i = n$. We use the symbol Π_n to denote the space of perfect phylogenies of size n sequences, and we use $\pi \in \Pi_n$ to denote a perfect phylogeny with n sequences.

A perfect phylogeny π is completely specified in a parenthetical notation, in which every leaf is represented by its label, every binary internal node is represented by $(\cdot, \cdot)$ and every multifurcating internal node is represented by $(\cdot, \ldots, \cdot)$. For example, the perfect phylogeny π_1 on the right in Figure 3B in parenthetical notation is $((2, 1), 2)$, indicating that there are two internal nodes, one merging leaves $(2, 1)$ and one merging $(2, 1)$ with 2.

The most extreme unresolved perfect phylogeny with n tips—the perfect phylogeny that is compatible with all ranked tree shapes with n tips—has two representations. It can be written as a star, in which the root has degree n and is the only internal node, that is, $\pi = (1, 1, \ldots, 1)$. It can also be written as a single node $\pi = (n)$. For our purposes, with mutations discarded, the star and single-node perfect phylogenies are indistinguishable, and they will be represented as a single-node perfect phylogeny. Details of the algorithm for generating the perfect phylogeny from binary molecular data can be found in Cappello et al. (2020a), which presents a slight modification to Gusfield's algorithm (Gusfield, 1991).

We say that a binary tree T is **compatible** with a perfect phylogeny π if the tree can be reduced to π by collapsing internal edges of T. The number of tree shapes, ranked or unranked, that are compatible with a perfect phylogeny gives the cardinality of the corresponding posterior sampling tree space in statistical inference from sequence data sets. Given a perfect phylogeny $\pi \in \Pi_n$, we are interested in calculating the number of compatible ranked tree shapes with n leaves and the number of compatible unranked tree shapes with n leaves.

2.4 Known enumerative results

In advance of our effort to count tree shapes compatible with a perfect phylogeny, we state some known enumerative results for the unconstrained spaces of ranked labeled tree shapes, unranked labeled tree shapes, ranked unlabeled tree shapes, and unranked unlabeled tree shapes (Steel, 2016).

Let $L_n = |\mathcal{T}_n^L|$ denote the cardinality of the space of ranked labeled trees with n leaves. Then

$$L_n = \prod_{i=2}^{n} \binom{i}{2} = \frac{n!(n-1)!}{2^{n-1}}. \tag{1}$$

The product is obtained by noting that for each decreasing i from n to 2, there are $\binom{i}{2}$ ways of merging two labeled branches. The sequence of values of L_n begins 1, 1, 3, 18, 180, 2700, 56700.

Let $X_n = |\mathcal{T}_n^X|$ denote the number of unranked labeled trees with n leaves. We have

$$X_n = (2n - 3)!! = \frac{(2n - 2)!}{2^{n-1}(n - 1)!}. \tag{2}$$

To generate trees in $\mathcal{T}_n^X$ from trees in $\mathcal{T}_{n-1}^X$, a pendant edge connected to the nth label can be placed along each of the $2n - 3$ edges of a tree with $n - 1$ leaves, including an edge above the root.

X_n is obtained as the solution to the recursion $X_n = (2n-3)X_{n-1}$, with $X_1 = 1$. The sequence of values of X_n begins $1, 1, 3, 15, 105, 945, 10395$.

The number of ranked tree shapes with n tips is the $(n-1)$-th Euler zigzag number (Stanley). Let $R_n = |\mathcal{T}_n^R|$ denote the number of ranked tree shapes with n leaves. We have the following recursion:

$$R_1 = 1, \ R_2 = 1,$$

$$R_{n+1} = \frac{1}{2} \sum_{k=0}^{n-1} \binom{n-1}{k} R_{k+1} R_{n-k}, \ n \geq 2. \tag{3}$$

The sequence of values of R_n begins $1, 1, 1, 2, 5, 16, 61$. For $n \geq 1$, if the tree has $n+1$ tips, and hence n interior nodes, then the root divides the tree into two ranked subtrees T_1^R and T_2^R, where T_1^R has k interior nodes, $0 \leq k \leq n-1$, and T_2^R has $n-1-k$ interior nodes. There are $\binom{n-1}{k}$ ways of interleaving the k and $n-1-k$ interior nodes of T_1^R and T_2^R, such that the relative orderings of the interior nodes of T_1^R and T_2^R are preserved in the interleaving. The number of possible ranked tree shapes with such a configuration is $\binom{n-1}{k} R_{k+1} R_{n-k}$. Summing over the possibilities for k from 0 to $n-1$, and acknowledging that the identity of T_1^R and T_2^R can be interchanged, we get eq. 3.

Let $S_n = |\mathcal{T}_n|$ denote the number of unranked tree shapes with n leaves. We have the following recursion:

$$S_1 = 1,$$

$$S_{2n-1} = \sum_{k=1}^{n-1} S_k S_{2n-1-k}, \ n \geq 2, \tag{4}$$

$$S_{2n} = \left(\sum_{k=1}^{n-1} S_k S_{2n-k} \right) + \frac{1}{2} S_n (S_n + 1), \ n \geq 1. \tag{5}$$

S_n is the nth Wedderburn-Etherington number (Harding, 1971). The sequence begins $1, 1, 1, 2, 3, 6, 11$. When the number of leaves is $2n-1$, the root divides the tree shape into two subtree shapes T_1 and T_2 with k and $2n-1-k$ leaves, for $k = 1, 2, \ldots, n-1$. When the number of leaves is even, the root divides the tree shape into subtree shapes with k and $2n-k$ leaves for $k = 1, 2, \ldots, n-1$ or two subtree shapes with n leaves; these tree shapes are indistinguishable in S_n cases and distinguishable in $\frac{1}{2} S_n (S_n - 1)$ cases.

3 Enumeration for binary perfect phylogenies

To count ranked and unranked tree shapes compatible with a perfect phylogeny, we first consider binary perfect phylogenies: those perfect phylogenies for which the outdegree of any node, traversing from root to tips, is either 0 (leaves or taxa) or 2 (internal nodes). We then consider multifurcating perfect phylogenies in Section 4.

3.1 Lattice structure of binary perfect phylogenies

The binary perfect phylogenies for a set of n tips possess a structure that will assist in enumerating binary ranked and unranked trees compatible with a set of sequences. In particular, we can make
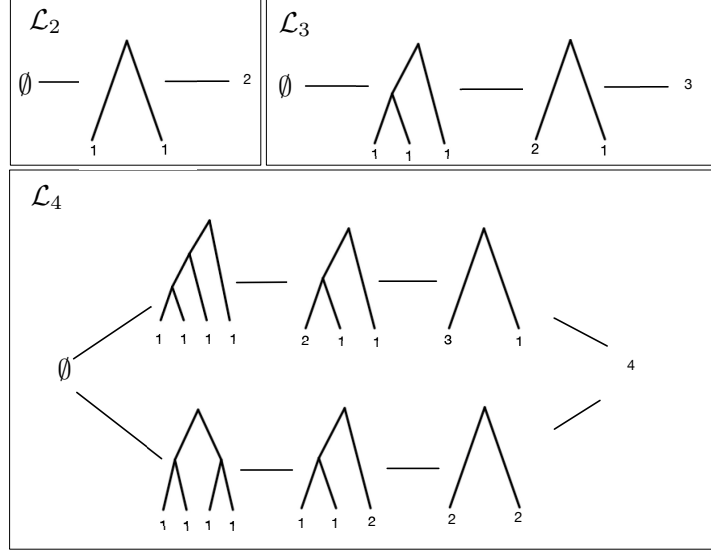
Figure 4: **Hasse diagrams of the lattices of binary perfect phylogenies with** $n = 2$, 3, **and** 4 **taxa.**

the set Π_n of all binary perfect phylogenies of $[n]$ into a **poset** by defining $\pi \leq \sigma$ if either σ is the same as π, or if σ can be obtained by sequentially collapsing pairs of pendant edges, or cherries, of π. We then say π is a **refinement** of σ. For example, $\pi = (2, 3)$ refines $\sigma = (5)$. We say that two binary perfect phylogenies in Π_n are **comparable** if they are equal or if one is a refinement of the other. An example of two perfect phylogenies that are not comparable is $\pi = (2, 3)$ and $\sigma = (4, 1)$.

Given two binary perfect phylogenies π_1 and π_2 in Π_n, their **meet**, denoted $\pi_1 \wedge \pi_2$, is the largest perfect phylogeny that refines both π_1 and π_2. Similarly, the **join** of two binary perfect phylogenies $\pi_1 \vee \pi_2$ is the smallest perfect phylogeny that is refined by both π_1 and π_2. Formal definitions of these notions appear in Definition 1.

Under the meet and join operations, we will see in Theorem 4 that the poset $\Pi_n \cup \{\emptyset\}$ is a **lattice** $\mathcal{L}_n = (\Pi_n \cup \{\emptyset\}, \wedge, \vee)$. As a lattice, $\mathcal{L}_n$ possesses a **Hasse diagram** with a minimal and a maximal element. The **maximal** element of $\mathcal{L}_n$ is the single node perfect phylogeny (n) and the **minimal** element is $\emptyset$. Figures 4 and 5 show the Hasse diagrams of $\mathcal{L}_2$, $\mathcal{L}_3$, $\mathcal{L}_4$, $\mathcal{L}_5$.

Definition 1. Binary perfect phylogeny operations. We define the binary perfect phylogeny symmetric operations $\wedge, \vee : (\cup_{n \geq 1} \Pi_n \cup \{\emptyset\}) \times (\cup_{n \geq 1} \Pi_n \cup \{\emptyset\}) \to (\cup_{n \geq 1} \Pi_n \cup \{\emptyset\})$, where Π_n is the space of binary perfect phylogenies of n leaves, as follows:

1. $\pi \wedge \emptyset = \emptyset$, for all $\pi \in \Pi_n$.

2. $\pi \vee \emptyset = \pi$, for all $\pi \in \Pi_n$.

3. $\pi \wedge (n) = \pi$, for all $\pi \in \Pi_n$.

4. $\pi \vee (n) = (n)$, for all $\pi \in \Pi_n$.

5. $\pi_1 \wedge \pi_2 = \emptyset$, for all $\pi_1 \in \Pi_{n_1}, \pi_2 \in \Pi_{n_2}$, with $n_1 \neq n_2$.

6. $\pi_1 \vee \pi_2 = \emptyset$, for all $\pi_1 \in \Pi_{n_1}, \pi_2 \in \Pi_{n_2}$, with $n_1 \neq n_2$.
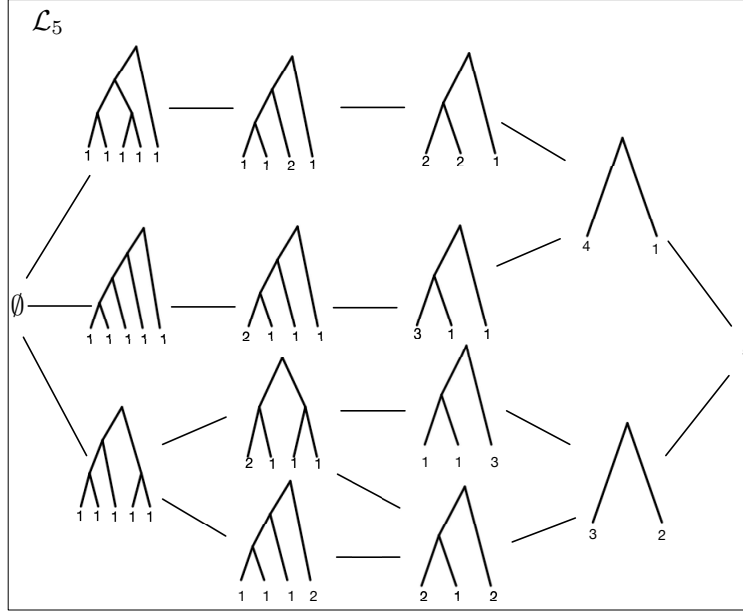
8

Figure 5: **Hasse diagram of the lattice of binary perfect phylogenies with $n = 5$ taxa.**

7. Let $\pi_1 = (n_1, n_2)$ and $\pi_2 = (n_3, n_4)$ be two perfect phylogenies in Π_n with $n_1 + n_2 = n_3 + n_4 = n$. Then

$$\pi_1 \vee \pi_2 = (n_1, n_2) \vee (n_3, n_4) = \begin{cases} (n_1, n_2) & \text{if } n_1 = n_3 \text{ or } n_1 = n_4 \\ (n) & \text{otherwise.} \end{cases}$$

8. For all π_1, π_2, π_3, π_4 with $(\pi_1, \pi_2) \in \Pi_n$ and $(\pi_3, \pi_4) \in \Pi_n$,

$$(\pi_1, \pi_2) \wedge (\pi_3, \pi_4) = (\pi_1 \wedge \pi_3, \pi_2 \wedge \pi_4) \vee (\pi_1 \wedge \pi_4, \pi_2 \wedge \pi_3),$$

with the convention that $(\pi, \emptyset) = \emptyset$. That is, the meet of two perfect phylogenies is the join of the two perfect phylogenies formed by merging two subtrees at the root. These four subtrees (two per newly formed perfect phylogeny) correspond to the meets of all pairs of subtrees, one from each of the original perfect phylogenies.

9. For all π_1, π_2, π_3, π_4 with $(\pi_1, \pi_2) \in \Pi_n$ and $(\pi_3, \pi_4) \in \Pi_n$, $\pi_i \in \Pi_{n_i}$ for $i = 1, 2, 3, 4$.

$$(\pi_1, \pi_2) \vee (\pi_3, \pi_4) = \begin{cases} (n) & \text{if } n_1 \neq n_3 \text{ and } n_1 \neq n_4 \\ (\pi_1, \pi_2 \vee \pi_4) & \text{if } \pi_1 = \pi_3 \\ (\pi_1, \pi_2 \vee \pi_3) & \text{if } \pi_1 = \pi_4 \\ (\pi_2, \pi_1 \vee \pi_4) & \text{if } \pi_2 = \pi_3 \\ (\pi_2, \pi_1 \vee \pi_3) & \text{if } \pi_2 = \pi_4 \\ (\pi_1 \vee \pi_3, \pi_2 \vee \pi_4) \wedge (\pi_1 \vee \pi_4, \pi_2 \vee \pi_3) & \text{otherwise,} \end{cases}$$

with the convention that $(\pi, \emptyset) = \emptyset$. That is, the join of two perfect phylogenies is the meet of the two perfect phylogenies formed by merging two subtrees at the root. These four subtrees

9

(two per newly formed perfect phylogeny) correspond to the joins of all pairs of subtrees, one from each of the original perfect phylogenies. In the particular case that the two original perfect phylogenies share one of the subtrees descending from the root, then the join of the two perfect phylogenies is the perfect phylogeny that merges, at the root, the shared subtree with the join of the two different subtrees, one from each of the original perfect phylogenies. In the case that no two pairs of subtrees, one from each of the original perfect phylogenies, have the same size, the join is the maximal single node perfect phylogeny (n).

10. For all $\pi_1, \pi_2, \pi_3 \in \Pi_n$,
$$\pi_1 \wedge (\pi_2 \vee \pi_3) = (\pi_1 \wedge \pi_2) \vee (\pi_1 \wedge \pi_3),$$
and
$$\pi_1 \vee (\pi_2 \wedge \pi_3) = (\pi_1 \vee \pi_2) \wedge (\pi_1 \vee \pi_3).$$

11. Let $\pi, \sigma \in \Pi_n$ be two perfect phylogenies that are not comparable. There exist unique $\gamma, \rho \in (\Pi_n \cup \{\emptyset\}) \setminus \{\pi, \sigma\}$ such that
$$\pi \wedge \sigma = \gamma, \quad \pi \vee \gamma = \pi, \quad \text{and} \quad \sigma \vee \gamma = \sigma,$$
and
$$\pi \vee \sigma = \rho, \quad \pi \wedge \rho = \pi, \quad \text{and} \quad \sigma \wedge \rho = \sigma.$$

Note that the meet and join operations are symmetric and that pairs (π_1, π_2) are unordered; for convenience, we have expanded expressions in parts 7 and 9 of the definition that could potentially be simplified using the symmetry.

We illustrate the operations in Definition 1 by considering a series of examples.

Example 2. Consider $\pi_1 = ((4, 2), 6)$ and $\pi_2 = ((3, 3), 6)$ depicted in Figure 6A. Their meet and join are given by:

$$
\begin{aligned}
((4, 2), 6) \wedge ((3, 3), 6) &= ((4, 2) \wedge (3, 3), 6 \wedge 6) \vee ((4, 2) \wedge 6, 6 \wedge (3, 3)) \text{ by Defn. 1 (8)} \\
&= (\emptyset, 6) \vee ((4, 2), (3, 3)) \text{ by Defn. 1 (3, 5, 8)} \\
&= \emptyset \vee ((4, 2), (3, 3)) \text{ by convention} \\
&= ((4, 2), (3, 3)) \text{ by Defn. 1 (2).}
\end{aligned}
$$

$$
\begin{aligned}
((4, 2), 6) \vee ((3, 3), 6) &= (6, (4, 2) \vee (3, 3)) \text{ by Defn. 1 (9)} \\
&= (6, 6) \text{ by Defn. 1 (7).}
\end{aligned}
$$

Example 3. For a more complex example, consider $\pi_1 = ((3, 1), 2), 6)$ and $\pi_2 = ((4, 2), 6)$ depicted in Figure 6B.

$$
\begin{aligned}
(((3, 1), 2), 6) \wedge ((4, 2), 6) &= (((3, 1), 2) \wedge (4, 2), 6 \wedge 6) \vee (((3, 1), 2) \wedge 6, 6 \wedge (4, 2)) \text{ by Defn. 1 (8)} \\
&= (((3, 1), 2) \wedge (4, 2), 6) \vee (((3, 1), 2), (4, 2)) \text{ by Defn. 1 (3)} \\
&= (((3, 1) \wedge 4, 2 \wedge 2), 6) \vee (((3, 1), 2), (4, 2)) \text{ by Defn. 1 (2, 5, 8)} \\
&= (((3, 1), 2), 6) \vee (((3, 1), 2), (4, 2)) \text{ by Defn. 1 (3)} \\
&= (((3, 1), 2), 6) \text{ by Defn. 1 (4, 9).}
\end{aligned}
$$

$$
\begin{aligned}
((3, 1), 2), 6) \vee ((4, 2), 6) &= (((3, 1), 2) \vee (4, 2), 6) \text{ by Defn. 1 (9)} \\
&= ((4, 2), 6) \text{ by Defn. 1 (4, 9).}
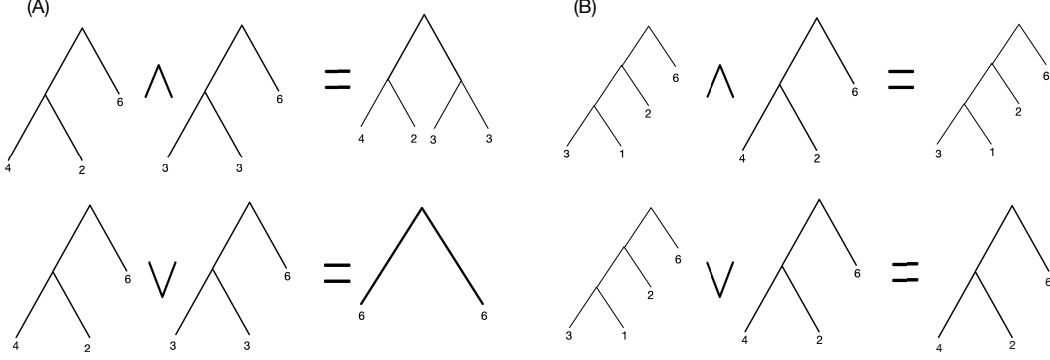\end{aligned}
$$

Figure 6: **Examples of perfect phylogeny operations. (A)** For perfect phylogenies $((4,2),6)$ and $((3,3),6)$, their meet is $((4,2),(3,3))$, and their join is $(6,6)$. **(B)** For perfect phylogenies $(((3,1),2),6)$ and $((4,2),6)$, their meet is $(((3,1),2),6)$ and their join is $((4,2),6)$.

To make use of the operations $\wedge$ and $\vee$ for counting binary ranked and unranked trees compatible with a perfect phylogeny, we need a theorem that shows that the two operations $\wedge$ and $\vee$ induce the same order. That is, we will show that $(\Pi_n \cup \{\emptyset\}, \wedge, \vee)$ is a lattice.

A *lattice* (Nation, 1998) is an algebra $\mathcal{L}(L, \wedge, \vee)$ satisfying, for all $x, y, z \in L$,

1. $x \wedge x = x$ and $x \vee x = x$,

2. $x \wedge y = y \wedge x$ and $x \vee y = y \vee x$,

3. $x \wedge (y \wedge z) = (x \wedge y) \wedge z$ and $x \vee (y \vee z) = (x \vee y) \vee z$,

4. $x \wedge (x \vee y) = x$ and $x \vee (x \wedge y) = x$.

In the Appendix, we verify these conditions for $(\Pi_n \cup \{\emptyset\}, \wedge, \vee)$, giving the following theorem.

Theorem 4. $(\Pi_n \cup \{\emptyset\}, \wedge, \vee)$ *is a lattice.*

3.2 Unranked unlabeled tree shapes compatible with a binary perfect phylogeny

With the lattice structure of the binary perfect phylogenies established, we are now equipped to calculate the number of compatible unranked unlabeled tree shapes with n leaves. Notice that an unranked unlabeled tree shape can be transformed into a perfect phylogeny with the same number of tips by assigning the count 1 to all leaves. We use $\mathcal{P}(T_n)$ to denote the perfect phylogeny with n tips that corresponds to the unranked unlabeled tree shape T_n.

Definition 5. Unranked unlabeled tree shape T_n compatible with a perfect phylogeny $\pi \in \Pi_n$. An unranked unlabeled tree shape with n leaves, T_n, is compatible with a perfect phylogeny $\pi \in \Pi_n$, if (1) a one-to-one correspondence exists between the k leaves of π with leaf counts $n_1, n_2, \ldots, n_k$ and k disjoint subtrees of T_n containing $n_1, n_2, \ldots, n_k$ leaves, respectively; and (2) $\mathcal{P}(T_n) \leq \pi$, that is, $\mathcal{P}(T_n)$ is a refinement of π.
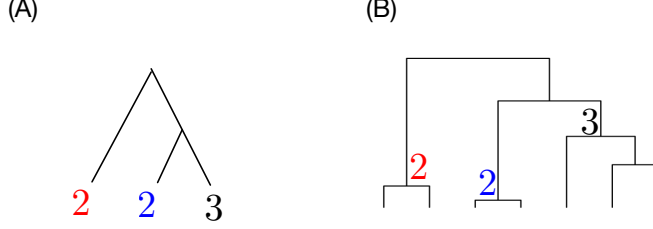
11

Figure 7: **Example of a tree shape compatible with a perfect phylogeny. (A)** A perfect phylogeny. **(B)** An unranked unlabeled tree shape that is compatible with the perfect phylogeny in (A). The numbers indicate the one-to-one correspondence described in Definition 5.

.

We use the symbol $\mathcal{G}_c(\pi) = \{T_n : T_n \rightsquigarrow \pi\}$ to denote the set of unranked unlabeled tree shapes compatible with a perfect phylogeny $\pi \in \Pi_n$. For a perfect phylogeny π consisting of a single leaf with leaf count n, the number of compatible unranked unlabeled tree shapes is simply the number of unranked unlabeled tre shapes of size n, or $|\mathcal{G}_c(\pi)| = S_n$. Figure 7 shows an example of an unranked unlabeled tree shape compatible with a perfect phylogeny of sample size 7.

Proposition 6. *For $n_1, n_2 \geq 1$, the number of unranked unlabeled tree shapes compatible with a cherry perfect phylogeny $(n_1, n_2) \in \Pi_n$ is*

$$|\mathcal{G}_c((n_1, n_2))| = \begin{cases} S_{n_1} S_{n_2} & \text{if } n_1 \neq n_2 \\ \frac{1}{2} S_{n_1}(S_{n_1} + 1) & \text{if } n_1 = n_2. \end{cases} \tag{6}$$

Proof. By Definition 5, an unranked unlabeled tree shape is compatible with the perfect phylogeny $\pi = (n_1, n_2)$ if it possesses two subtrees, one with n_1 leaf descendants and another with n_2 leaf descendants. Decomposing an unranked unlabeled tree shape at its root, the number of shapes with this property is $S_{n_1} S_{n_2}$ for $n_1 \neq n_2$ and $\frac{1}{2} S_{n_1}(S_{n_1} + 1)$ for $n_1 = n_2$. $\square$

Proposition 7. *For $n_1, n_2 \geq 1$ and $\pi_1 \in \Pi_{n_1}$, $\pi_2 \in \Pi_{n_2}$, the number of unranked unlabeled tree shapes compatible with a binary perfect phylogeny $\pi = (\pi_1, \pi_2) \in \Pi_n$ is*

$$|\mathcal{G}_c((\pi_1, \pi_2))| = \begin{cases} |\mathcal{G}_c(\pi_1)| \, |\mathcal{G}_c(\pi_2)| - \frac{1}{2}|\mathcal{G}_c(\pi_1 \wedge \pi_2)| \, (|\mathcal{G}_c(\pi_1 \wedge \pi_2)| - 1) & \text{if } \pi_1 \wedge \pi_2 \neq \emptyset \\ |\mathcal{G}_c(\pi_1)| \, |\mathcal{G}_c(\pi_2)| & \text{if } \pi_1 \wedge \pi_2 = \emptyset. \end{cases} \tag{7}$$

Proof. If $\pi_1 \wedge \pi_2 = \emptyset$, then no tree shapes are compatible with both π_1 and π_2. Hence, the number of tree shapes compatible with (π_1, π_2) is simply the product of the number of tree shapes compatible with π_1 and the number of tree shapes compatible with π_2.

If $\pi_1 \wedge \pi_2 \neq \emptyset$, then certain tree shapes can be compatible with both π_1 and π_2, i.e., compatible with $\pi_1 \wedge \pi_2$. We sum four quantities. (1) Consider the set of tree shapes compatible with both perfect phylogenies π_1 and π_2. They can either be assigned the same tree shape, in $|\mathcal{G}_c(\pi_1 \wedge \pi_2)|$ ways, or they can be assigned different tree shapes, in $\frac{1}{2}(|\mathcal{G}_c(\pi_1 \wedge \pi_2)|^2 - |\mathcal{G}_c(\pi_1 \wedge \pi_2)|)$ ways, resulting in $\frac{1}{2}|\mathcal{G}_c(\pi_1 \wedge \pi_2)|(|\mathcal{G}_c(\pi_1 \wedge \pi_2)| + 1)$ tree shapes. (2) If π_2 is a refinement of π_1 and $\pi_1 \neq \pi_2$, then there are $|\mathcal{G}_c(\pi_1 \wedge \pi_2)|(|\mathcal{G}_c(\pi_1)| - |\mathcal{G}_c(\pi_1 \wedge \pi_2)|)$ tree shapes. (3) Similarly, if π_1 is a refinement of π_2 and $\pi_1 \neq \pi_2$,

12

then there are $|\mathcal{G}_c(\pi_1 \wedge \pi_2)|(|\mathcal{G}_c(\pi_2)| - |\mathcal{G}_c(\pi_1 \wedge \pi_2)|)$. (4) If π_1 and π_2 are not comparable, that is, if neither is a refinement of the other, then there are $(|\mathcal{G}_c(\pi_1)| - |\mathcal{G}_c(\pi_1 \wedge \pi_2)|)(|\mathcal{G}_c(\pi_2)| - |\mathcal{G}_c(\pi_1 \wedge \pi_2)|)$ tree shapes. Scenarios (2), (3), and (4) are mutually exclusive, and only one of the quantities in (2), (3), and (4) is nonzero; summing the four quantities gives the result. $\qquad \square$

Propositions 6 and 7 provide a recursive formula for calculating the number of tree shapes compatible with a binary perfect phylogeny. For example, examining Figure 6A, the number of tree shapes compatible with $(4, 2)$ is $S_4 S_2 = 2$, and the number of tree shapes compatible with $((4, 2), 6)$ is $|\mathcal{G}_c(4, 2)| |\mathcal{G}_c(6)| - \frac{1}{2}|\mathcal{G}_c(4, 2)|(|\mathcal{G}_c(4, 2)| - 1) = (2)(6) - \frac{1}{2}(2)(1) = 11$. Table 1 shows the number of tree shapes compatible with certain perfect phylogenies of sample size 10.

3.3 Ranked unlabeled tree shapes compatible with a binary perfect phylogeny

Next, for a binary perfect phylogeny, we compute the number of compatible ranked unlabeled tree shapes with n leaves.

Definition 8. Ranked unlabeled tree shape T_n^R compatible with a perfect phylogeny $\pi \in \Pi_n$. A ranked unlabeled tree shape with n leaves, T_n^R, is compatible with a perfect phylogeny $\pi \in \Pi_n$ if the unranked unlabeled tree shape T_n obtained by removing the ranking from T_n^R is compatible with π.

Proposition 9. *For $n_1, n_2 \geq 1$, the number of ranked unlabeled tree shapes compatible with a cherry perfect phylogeny $(n_1, n_2) \in \Pi_n$ is*

$$|\mathcal{G}_c^T((n_1, n_2))| = \begin{cases} \binom{n_1+n_2-2}{n_1-1} R_{n_1} R_{n_2} & \text{if } n_1 \neq n_2 \\ \frac{1}{2}\binom{2n_1-2}{n_1-1} R_{n_1}^2 & \text{if } n_1 = n_2. \end{cases} \tag{8}$$

Proof. By Definition 8, a ranked unlabeled tree shape T^R is compatible with the perfect phylogeny $\pi = (n_1, n_2)$ if the associated unranked unlabeled tree shape T obtained by removing the ranking of T^R is compatible with π. By Definition 5, the unranked unlabeled tree shape T is compatible with the perfect phylogeny $\pi = (n_1, n_2)$ if it possesses two subtrees, one with n_1 leaf descendants and another with n_2 leaf descendants.

We decompose a ranked unlabeled tree at its root into subtrees of size n_1 and n_2. If $n_1 \neq n_2$, then the $n_1 - 1$ interior nodes of the subtree with n_1 leaves and the $n_2 - 1$ interior nodes of the subtree with n_2 leaves can be interleaved in $\binom{n_1+n_2-2}{n_1-1}$ ways. If $n_1 = n_2$, then the two ranked subtrees can be the same in R_{n_1} ways, each with $\frac{1}{2}\binom{2n_1-2}{n_1-1}$ ways of interleaving the two ranked unlabeled subtrees; the two ranked subtrees can differ in $\frac{1}{2}(R_{n_1}^2 - R_{n_1})$ ways, each with $\binom{2n_1-2}{n_1-1}$ ways of interleaving the subtrees. $\qquad \square$

Proposition 10. *For $n_1, n_2 \geq 1$ and $\pi_1 \in \Pi_{n_1}, \pi_2 \in \Pi_{n_2}$, the number of ranked unlabeled tree shapes compatible with a binary perfect phylogeny $\pi = (\pi_1, \pi_2) \in \Pi_n$ is*

$$|\mathcal{G}_c^T((\pi_1, \pi_2))| = \begin{cases} \binom{2n_1-2}{n_1-1}(|\mathcal{G}_c^T(\pi_1)| |\mathcal{G}_c^T(\pi_2)| - \frac{1}{2}|\mathcal{G}_c^T(\pi_1 \wedge \pi_2)|^2) & \text{if } \pi_1 \wedge \pi_2 \neq \emptyset \\ \binom{n_1+n_2-2}{n_1-1}|\mathcal{G}_c^T(\pi_1)| |\mathcal{G}_c^T(\pi_2)| & \text{if } \pi_1 \wedge \pi_2 = \emptyset. \end{cases} \tag{9}$$

Proof. If $\pi_1 \wedge \pi_2 = \emptyset$, then the number of ranked tree shapes compatible with (π_1, π_2) is simply the product of the number of ranked tree shapes compatible with π_1, the number of ranked tree shapes compatible with π_2, and the number of ways of interleaving their rankings.

If $\pi_1 \wedge \pi_2 \neq \emptyset$, then certain ranked tree shapes can be compatible with both π_1 and π_2, i.e., compatible with $\pi_1 \wedge \pi_2$. We therefore have three cases: the two perfect phylogenies are the same, one is a refinement of the other (two possible ways), or neither is a refinement of the other. The cardinalities in these cases are $\frac{1}{2}|\mathcal{G}_c^T(\pi_1 \wedge \pi_2)|^2$, $|\mathcal{G}_c^T(\pi_1 \wedge \pi_2)|(|\mathcal{G}_c^T(\pi_2)| - |\mathcal{G}_c^T(\pi_1 \wedge \pi_2)|) + |\mathcal{G}_c^T(\pi_1 \wedge \pi_2)|(|\mathcal{G}_c^T(\pi_1)| - |\mathcal{G}_c^T(\pi_1 \wedge \pi_2)|)$, and $(|\mathcal{G}_c^T(\pi_1)| - |\mathcal{G}_c^T(\pi_1 \wedge \pi_2)|)(|\mathcal{G}_c^T(\pi_2)| - |\mathcal{G}_c^T(\pi_1 \wedge \pi_2)|)$, respectively, all multiplied by the possible number of interleavings of the rankings $\binom{2n_1-2}{n_1-1}$. $\square$

Propositions 9 and 10 provide a recursive formula for calculating the number of ranked tree shapes compatible with a binary perfect phylogeny. For Figure 6A, the number of ranked tree shapes compatible with $(4, 2)$ is $(4)(2) = 8$, and the number of ranked tree shapes compatible with $((4, 2), 6)$ is $\binom{10}{5}(|\mathcal{G}_c^T(4, 2)| |\mathcal{G}_c^T(6)| - \frac{1}{2}|\mathcal{G}_c^T(4, 2)|^2) = \binom{10}{5}[(8)(16) - \frac{1}{2}(8)^2] = 24,192$.

Table 1 shows the number of ranked unlabeled tree shapes compatible with some of the perfect phylogenies of sample size 10. We can observe that these numbers exceed corresponding numbers of unranked unlabeled tree shapes compatible with the perfect phylogenies, just as the numbers of ranked unlabeled tree shapes exceed the numbers of unranked unlabeled tree shapes (Section 2.4).

For the ranked unlabeled tree shapes compatible with a binary perfect phylogeny, we can examine the asymptotic growth of the number of compatible ranked unlabeled tree shapes in particular families of binary perfect phylogenies. For a fixed integer value $x \geq 1$, consider the family of binary perfect phylogenies $B_x(n) = (x, n - x)$ as n increases. These are cherry phylogenies with labels x and $n - x$ at their two leaves. Let $b_x(n)$ be the number of ranked unlabeled tree shapes compatible with $B_x(n)$. Among the integer sequences $b_1(n)$, $b_2(n)$, $b_3(n)$, ..., the next proposition shows that $b_2(n)$ has the fastest asymptotic growth. In other words, as n grows large, the value of x for which the number of ranked unlabeled tree shapes compatible with the perfect phylogeny $B_x(n)$ is asymptotically largest is $x = 2$.

Proposition 11. *Among the integer sequences $b_1(n)$, $b_2(n)$, $b_3(n)$, ..., the sequence $b_2(n)$ has the fastest asymptotic growth.*

Proof. For a fixed integer value $x \geq 0$, let $\beta_x = (x + 1, n - x + 1)$ be a binary perfect phylogeny with two leaves, labeled by $x + 1$ (say to the left of the root) and $n - x + 1$ (to the right of the root). The set of ranked unlabeled tree shapes compatible with β_x corresponds to the set of ranked unlabeled tree shapes with $n + 1$ internal nodes ($n + 2$ leaves), x internal nodes for the left root subtree, and $n - x$ internal nodes for the right root subtree.

We consider an increasing sequence of values of n. Supposing $n > 2x$ so that the root subtrees of β_x cannot have the same sample size, we apply Proposition 10, finding that the number of ranked unlabeled tree shapes compatible with β_x is

$$\binom{n}{x} e_x e_{n-x}, \tag{10}$$

where e_i is the number of ranked unlabeled tree shapes with i internal nodes. Following eq. 3, the integer e_i is the ith Euler number, $e_i = R_{i+1}$.

The exponential generating function of the sequence (e_i) is (Brent and Harvey, 2013)

$$\sum_{i=0}^{\infty} \frac{e_i z^i}{i!} = \sec(z) + \tan(z). \tag{11}$$

14

We can write the ratio $q_i = \frac{e_i}{i!}$ as (Flajolet and Sedgewick 2009, p. 269; Brent and Harvey, 2013)

$$q_i = \begin{cases} 2 \left(\frac{2}{\pi}\right)^{i+1} \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)^{i+1}}, & \text{if } i \text{ is even} \\ 2 \left[\left(\frac{2}{\pi}\right)^{i+1} - \left(\frac{1}{\pi}\right)^{i+1}\right] \sum_{k=1}^{\infty} \frac{1}{k^{i+1}}, & \text{if } i \text{ is odd.} \end{cases} \tag{12}$$

As i becomes large, by applying singularity analysis to eq. 11, or by computing directly from eq. 12, we have the asymptotic relation

$$q_i \sim 2 \left(\frac{2}{\pi}\right)^{i+1}. \tag{13}$$

With $q_x = e_x/x!$, we rewrite eq. 10 as $n! \, q_x q_{n-x}$. Letting $n \to \infty$ for a fixed x, we can use eq. 12 to rewrite q_x, and because x is constant as n grows, we can use eq. 13 for the asymptotic value of q_{n-x}. Hence, for increasing values of n, the number of ranked tree shapes compatible with the perfect phylogeny β_x behaves asymptotically like the product of $n!$ and

$$q_x q_{n-x} \sim 4 \left(\frac{2}{\pi}\right)^{n+2} c_x, \tag{14}$$

where

$$c_x = \begin{cases} \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)^{x+1}}, & \text{if } x \text{ is even} \\ \left(1 - \frac{1}{2^{x+1}}\right) \sum_{k=1}^{\infty} \frac{1}{k^{x+1}}, & \text{if } x \text{ is odd.} \end{cases} \tag{15}$$

Note that $\zeta(s) = \sum_{k=1}^{\infty} \frac{1}{k^s}$ is the Riemann zeta function. If x is even, then

$$c_x = 1 + \left(-\frac{1}{3^{x+1}} + \frac{1}{5^{x+1}}\right) + \left(-\frac{1}{7^{x+1}} + \frac{1}{9^{x+1}}\right) + \dots \le 1.$$

Among odd values of x, we have $c_1 = \frac{3}{4} \zeta(2) = \pi^2/8 \approx 1.2337$ for $x = 1$. For odd $x \ge 3$, we have

$$c_x < \zeta(x+1) \le \zeta(3) \approx 1.2021 < c_1.$$

Hence, $c_1 > 1$ exceeds c_x both for even x and for all odd $x \ge 3$.

Because c_x has its maximum at $x = 1$, from eq. 14, we conclude that the product $q_x q_{n-x}$ grows asymptotically fastest for $x = 1$. In particular, as $n \to \infty$, the value of x for which the binary perfect phylogeny β_x has the largest number of compatible ranked unlabeled tree shapes is $x = 1$—that is, when $\beta_x = \beta_1 = (2, n)$. □

In Table 1, we can observe an example of Proposition 11. The value of $b_2(10)$, or 2176, exceeds the values of $b_x(10)$ for all other values of x (with the trivial exception that $b_2(10) = b_8(10)$). The asymptotic approximation from eq. 14 gives

$$b_2(n) \sim 2 \left(\frac{2}{\pi}\right)^{n-2} (n-2)!,$$

which, for $n = 10$, yields $b_2(10) \approx 2175.66$.

3.4 Ranked labeled tree shapes compatible with a labeled binary perfect phylogeny

Propositions 6, 7, 9 and 10 provide recursive formulas for enumerating unranked unlabeled tree shapes and ranked unlabeled tree shapes compatible with a binary perfect phylogeny. In these cases, a perfect phylogeny representation does not use individual sequence labels; the labels of the tips of the perfect phylogeny are simply counts of numbers of sequences. We now consider **labeled perfect phylogenies** that partition the set of labeled individual sequences. We still use the parenthetical notation described in Section 2.3 to denote a labeled perfect phylogeny, for example $\pi = (2, 3)$, however, it must be understood that this labeled perfect phylogeny partitions the sampled sequences into two different sets of labeled sequences.

Consider $\{x_1, x_2\}$ and $\{x_3, x_4, x_5\}$ in the perfect phylogeny of Figure 8B. We are now interested in calculating the number of ranked labeled tree shapes compatible with a labeled binary perfect phylogeny. Figure 8C shows all the ranked labeled tree shapes compatible with the labeled perfect phylogeny. For ranked labeled tree shapes, the enumeration follows a simple recursive expression.

Definition 12. Ranked labeled tree shape T_n^L compatible with a labeled perfect phylogeny $\pi \in \Pi_n^L$. A ranked labeled tree shape with n leaves, T_n^L, is compatible with a perfect phylogeny $\pi \in \Pi_n^L$ if the unranked unlabeled tree shape T_n obtained by removing the ranks and the labels from T_n^L is compatible with π and the one-to-one correspondence between the k leaves of π and the k disjoint subtrees of T_n^L correspond to the same partition of the individual sequences.

Proposition 13. *For $n_1, n_2 \geq 1$ and $\pi_1 \in \Pi_{n_1}^L, \pi_2 \in \Pi_{n_2}^L$ the number of ranked labeled tree shapes compatible with a labeled binary perfect phylogeny $\pi = (\pi_1, \pi_2)$ is*

$$|\mathcal{G}_c^L(\pi)| = \binom{n_1 + n_2 - 2}{n_1 - 1} |\mathcal{G}_c^L(\pi_1)| \, |\mathcal{G}_c^L(\pi_2)|. \tag{16}$$

Proof. We can count the number of ranked labeled tree shapes by dividing π at the root into two subtrees, one with n_1 leaves and perfect phylogeny π_1, and the other with n_2 leaves and perfect phylogeny π_2, both partitioning the sampled sequences. The number of such trees is the product of the numbers of ranked labeled trees for the two subtrees and the number of ways of interleaving the internal nodes of the two subtrees. In this case, the two perfect phylogenies π_1 and π_2 can never be identical because they correspond to different sets of sequences. $\square$

Counts for the number of ranked labeled tree shapes for some of the perfect phylogenies of 10 samples (with an arbitrary labeling) appear in Table 1. Given a perfect phylogeny in the table, we can observe that the number of ranked labeled tree shapes far exceeds the number of ranked unlabeled tree shapes.

Continuing with $((4,2),6)$, the number of ranked labeled tree shapes compatible with this (arbitrarily labeled) perfect phylogeny is $\binom{10}{5}|\mathcal{G}_c^L((4,2))| \, |\mathcal{G}_c^L((6))| = \binom{10}{5}\binom{4}{3} |\mathcal{G}_c^L((4))| \, |\mathcal{G}_c^L((2))| \, |\mathcal{G}_c^L((6))| = \binom{10}{5}\binom{4}{3} L_4 L_2 L_6 = 252 \times 4 \times 18 \times 1 \times 2700 = 48,988,800$.

We can obtain a result analogous to Proposition 11; we characterize, for binary labeled perfect phylogenies $B_x(n) = (x, n - x)$, the one compatible with the largest number of ranked labeled tree shapes. Let $b_x'(n)$ denote the number of ranked labeled tree shapes compatible with $B_x(n)$.

Proposition 14. *Fix $n \geq 2$. Among the values $b_1'(n), b_2'(n), \ldots, b_{\lfloor \frac{n}{2} \rfloor}'(n)$, the largest is $b_1'(n)$.*
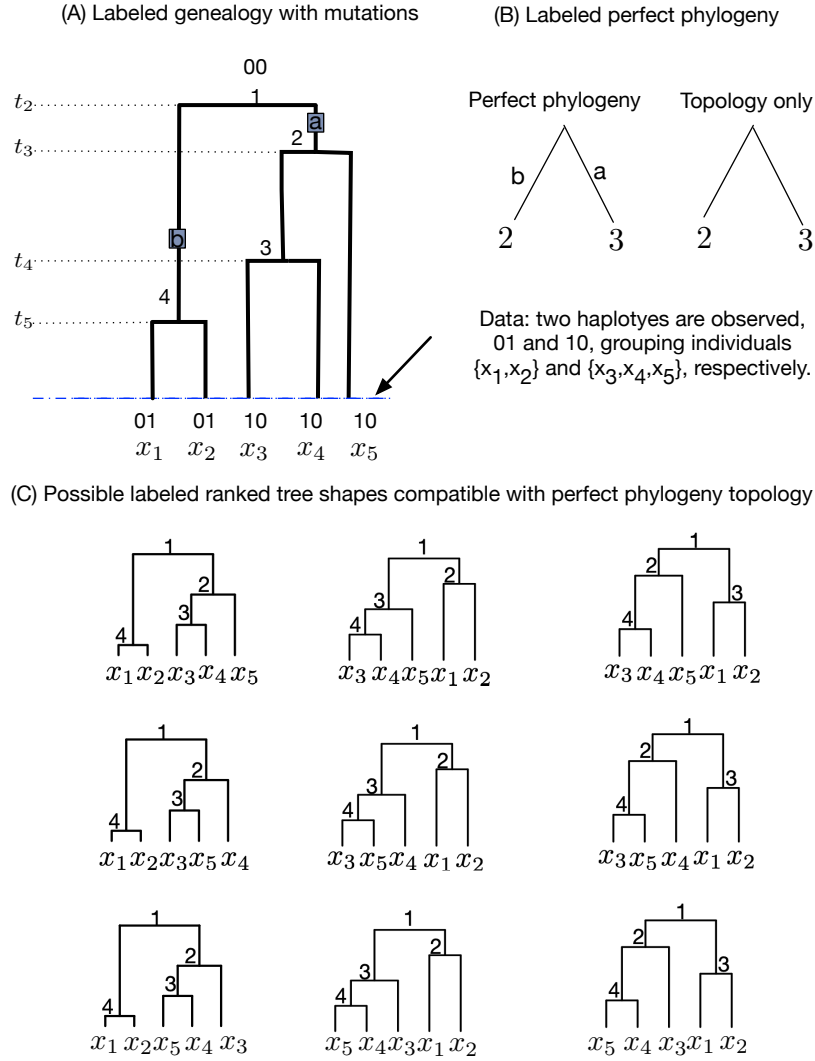
Figure 8: **Coalescent and infinitely-many-sites generative model of binary molecular data.** **(A)** A genealogy of 5 individuals, with 2 superimposed mutations depicted as gray squares. The root is labeled by the ancestral type 00, and the leaves are labeled by the genetic type at each of three mutated sites. The first two leaves from left to right are labeled 01 because one mutation occurs in their path to the root. The third, fourth and fifth individuals have one mutation in their path to the root and are labeled 10. The order and label of the mutations is unimportant; however, individual labels x_1, x_2, x_3, x_4, x_5 are important. For ease of exposition, we label the mutations a, b. The first site corresponds to mutation a, and the second to b. **(B)** Left, a labeled perfect phylogeny representation of the observed data at the tips of (A). Data consist of 2 unique haplotypes 01 and 10, with frequencies 2 and 3, respectively. The corresponding frequencies are the labels of tips of the perfect phylogeny; however, it is understood that the two leaves correspond to $\{x_1, x_2\}$ and $\{x_3, x_4, x_5\}$ respectively. Right, perfect phylogeny topology obtained by removing the edge labels of the perfect phylogeny. **(C)** The nine ranked labeled tree shapes compatible with the labeled perfect phylogeny topology in (B). Note that in (C), if we ignore the branching order and drop the internal node labels, in each row, the three trees are equivalent—so that each row corresponds to one of the three *unranked* labeled tree shapes compatible with the labeled perfect phylogeny topology in (B).

17

Proof. Applying Proposition 13 , we have $b'_x(n) = \binom{n-2}{x-1} L_x L_{n-x}$. Simplifying with eq. 1, we obtain $b'_x(n) = [n!\,(n-2)!/2^{n-2}]\binom{n}{x}^{-1}$. As it is quickly verified that the binomial coefficients $\binom{n}{x}$ increase monotonically from $x = 1$ to $x = \lfloor \frac{n}{2} \rfloor$, b'_x decreases monotonically from $x = 1$ to $x = \lfloor \frac{n}{2} \rfloor$. $\qquad\square$

An example of Proposition 14 is visible in Table 1, in which $b'_1(10) = 57,153,600$ exceeds $b'_2(10)$, $b'_3(10)$, $b'_4(10)$, and $b'_5(10)$.

3.5 Unranked labeled tree shapes compatible with a labeled binary perfect phylogeny

Continuing with the labeled perfect phylogenies from Section 3.4, we now count the unranked labeled binary perfect phylogenies compatible with a labeled binary perfect phylogeny.

Consider $\{x_1, x_2\}$ and $\{x_3, x_4, x_5\}$ in the perfect phylogeny of Figure 8B. We calculate the number of unranked labeled tree shapes compatible with a labeled binary perfect phylogeny. Each row of Figure 8C corresponds to one of the unranked labeled tree shapes compatible with the labeled perfect phylogeny.

Definition 15. Ranked labeled tree shape T_n^X compatible with a labeled perfect phylogeny $\pi \in \Pi_n^L$. An unranked labeled tree shape with n leaves, T_n^X, is compatible with a perfect phylogeny $\pi \in \Pi_n^L$ if the unranked unlabeled tree shape T_n obtained by removing the labels from T_n^X is compatible with π and the one-to-one correspondence between the k leaves of π and the k disjoint subtrees of T_n^X correspond to the same partition of the individual sequences.

Proposition 16. *For $n_1, n_2 \geq 1$ and $\pi_1 \in \Pi_{n_1}^L, \pi_2 \in \Pi_{n_2}^L$, the number of ranked labeled tree shapes compatible with a labeled binary perfect phylogeny $\pi = (\pi_1, \pi_2)$ is*

$$|\mathcal{G}_c^X(\pi)| = |\mathcal{G}_c^X(\pi_1)|\,|\mathcal{G}_c^X(\pi_2)|. \tag{17}$$

Proof. We divide π at the root into two subtrees, one with n_1 leaves and perfect phylogeny π_1, and the other with n_2 leaves and perfect phylogeny π_2. The subtrees must partition the sampled sequences in the same way as π. The number of such trees is the simply product of the numbers of unranked labeled trees for the two subtrees. As in Proposition 13, perfect phylogenies π_1 and π_2 are not identical because they correspond to different sets of sequences; with the ranking dropped, unlike in Proposition 13, we need not consider the number of ways of interleaving the internal nodes of the two subtrees. $\qquad\square$

For some of the perfect phylogenies of 10 samples (with an arbitrary labeling), counts for the number of unranked labeled tree shapes appear in Table 1. The number of unranked labeled tree shapes far exceeds the number of unranked unlabeled tree shapes, and it generally exceeds the number of ranked unlabeled tree shapes.

For the example $((4,2),6)$, the number of unranked labeled tree shapes compatible with this (arbitrarily labeled) perfect phylogeny is $|\mathcal{G}_c^X((4,2))|\,|\mathcal{G}_c^X((6))| = |\mathcal{G}_c^X((4))|\,|\mathcal{G}_c^X((2))|\,|\mathcal{G}_c^X((6))| = X_4 X_2 X_6 = 15 \times 1 \times 945 = 14,175$.

For binary labeled perfect phylogenies $B_x(n) = (x, n-x)$, the one compatible with the largest number of unranked labeled tree shapes follows the result of Proposition 14. Let $b''_x(n)$ denote the number of unranked labeled tree shapes compatible with $B_x(n)$.

18

Proposition 17. *Fix $n \geq 2$. Among the values $b_1''(n), b_2''(n), \ldots, b_{\lfloor \frac{n}{2} \rfloor}''(n)$, the largest is $b_1''(n)$.*

Proof. Applying Proposition 16 , we have $b_x''(n) = X_x X_{n-x}$ for $1 \leq x \leq \lfloor \frac{n}{2} \rfloor$. Simplifying with eq. 2, we obtain

$$b_x''(n) = \frac{(n-2)!}{2^{n-2}} \frac{\binom{2x-2}{x-1}\binom{2n-2x-2}{n-x-1}}{\binom{n-2}{x-1}}.$$

Then $b_{x+1}''(n)/b_x''(n) = \frac{2x-1}{2n-2x-3} \leq 1$ for $1 \leq x \leq \frac{n-1}{2}$, with equality requiring $x = \frac{n-1}{2}$, so that $b_x''(n)$ monotonically decreases from $x = 1$ to $x = \lfloor \frac{n}{2} \rfloor$. $\qquad \square$

In Table 1, we observe that as in Proposition 17, $b_1''(10) = 2{,}027{,}025$ exceeds $b_2''(10)$, $b_3''(10)$, $b_4''(10)$, and $b_5''(10)$.

4 Enumeration for multifurcating perfect phylogenies

Recall that perfect phylogenies need not be strictly binary, and that nodes can have more than two descendants. To complete the description of the numbers of trees of various types that are compatible with a perfect phylogeny, we must consider multifurcating perfect phylogenies. We proceed by reducing the multifurcating case to the binary case that has already been solved.

We now consider a **multifurcating perfect phylogeny** that consists of a single internal node subtending k leaves with labels $n_1, n_2, \ldots, n_k$. An example is depicted in Figure 9. Because multiple leaves can each correspond to groups with the same number of samples, so that the same numerical label can be assigned to many of those leaves, it is convenient to denote the vector of unique labels by $\mathbf{a} = (a_1, a_2, \ldots, a_s)$ and the corresponding vector of their multiplicities by $\mathbf{m} = (m_1, m_2, \ldots, m_s)$, where m_j denotes the number of leaves with label a_j, $1 \leq j \leq s \leq k$. In the example of Figure 9, $\mathbf{a} = (2, 3)$ and $\mathbf{m} = (2, 2)$, as two leaves ($m_1 = 2$) have label 2 ($a_1 = 2$) and two leaves ($m_2 = 3$) have label 3 ($a_2 = 3$).

We extend the notion of the binary perfect phylogeny poset to the multifurcating case. We define $\pi \leq \sigma$ for two multifurcating perfect phylogenies if σ can be obtained by sequentially collapsing pairs of pendant edges of π. Given two multifurcating perfect phylogenies π_1 and π_2, their meet $\pi_1 \wedge \pi_2$ is the largest multifurcating perfect phylogeny that refines both π_1 and π_2. For example, the meet between $\pi_1 = (1, 2, 3, (2, 2))$ and $\pi_2 = (1, 2, 2, (2, 3))$ is given by:

$$(1, 2, 3, (2, 2)) \wedge (1, 2, 2, (2, 3)) = (1, (2, 2), (2, 3)).$$

Similarly, their join is the smallest multifurcating perfect phylogeny $\pi_1 \vee \pi_2$ for which both π_1 and π_2 are refinements:

$$(1, 2, 3, (2, 2)) \vee (1, 2, 2, (2, 3)) = (1, 2, 2, 2, 3).$$

The lattice structure enables us to count the number of ranked unlabeled tree shapes compatible with a multifurcating perfect phylogeny $\pi = (n_1, n_2, \ldots, n_k)$. We use a recursive inclusion-exclusion principle with label vector $\mathbf{a}$ and multiplicities $\mathbf{m}$. The key idea is to decompose the computation into a sum over all possible binary perfect phylogenies, applying Propositions 9 and 10 to each binary perfect phylogeny. To recursively generate all possible binary perfect phylogenies from π,

we define the operator $\mathcal{B}_{i,j}(\pi)$ that collapses two leaves with labels a_i and a_j in π. For example $\mathcal{B}_{2,3}(2,2,3,4) = ((2,3),2,4)$. If $\sum_{i=1}^{s} m_i > 2$, then

$$
|\mathcal{G}_c(\pi)| = \underbrace{\sum_{i=1}^{s} |\mathcal{G}_c(\mathcal{B}_{i,i}(\pi))|\, 1_{m_i>1}}_{\substack{\text{collapsing two pendant edges} \\ \text{with the same leaf values}}} + \underbrace{\sum_{i=1}^{s-1} \sum_{j=i+1}^{s} |\mathcal{G}_c(\mathcal{B}_{i,j}(\pi))|\, 1_{m_i>0}\, 1_{m_j>0}}_{\substack{\text{collapsing two pendant edges} \\ \text{with different leaf values}}}
$$

$$
- \underbrace{\sum_{i=1}^{s-1} \sum_{j=i+1}^{s} |\mathcal{G}_c(\mathcal{B}_{i,i}(\pi) \wedge \mathcal{B}_{j,j}(\pi))|\, 1_{m_i>1}\, 1_{m_j>1}}_{\substack{\text{collapsing all pairs containing two distinct pairs of pendant edges,} \\ \text{each pair with the same leaf values}}}
$$

$$
- \underbrace{\sum_{i=1}^{s-1} \sum_{j=i+1}^{s} \sum_{\substack{k=1 \\ k\neq i, k\neq j}}^{s} |\mathcal{G}_c(\mathcal{B}_{i,j}(\pi) \wedge \mathcal{B}_{k,k}(\pi))|\, 1_{m_i>0}\, 1_{m_j>0}\, 1_{m_k>1}}_{\substack{\text{collapsing a pair of edges with different leaf values} \\ \text{and collapsing a pair of edges with the same leaf values}}}
$$

$$
- \underbrace{\sum_{i=1}^{s-1} \sum_{j=i+1}^{s} \sum_{\substack{k=1 \\ k\neq i, k\neq j}}^{s-1} \sum_{\substack{\ell=k+1 \\ \ell\neq i, \ell\neq j}}^{s} |\mathcal{G}_c(\mathcal{B}_{i,j}(\pi) \wedge \mathcal{B}_{k,\ell}(\pi))|\, 1_{m_i>0}\, 1_{m_j>0}\, 1_{m_k>0}\, 1_{m_\ell>0}}_{\substack{\text{collapsing two different pairs of pendant edges,} \\ \text{each pair with different leaf values}}}. \qquad (18)
$$

To interpret eq. 18 as an inclusion-exclusion formula, notice that the first two sums that are added on the right-hand side of eq. 18 correspond to enumerations of single events (so that the sum is analogous to a union $\cup A_i$), and the following three sums that are subtracted correspond to intersections of pairs of these events (analogous to intersections $A_i \cap A_j$).

Eq. 18 provides a recursive approach for counting the number of ranked unlabeled tree shapes compatible with a multifurcating perfect phylogeny by expressing the calculation in terms of binary perfect phylogenies. The recursive application of the equation proceeds until all terms reach $\sum_{i=1}^{s} m_i = 2$, when the binary perfect phylogenies are reached.

Example 18. The number of ranked unlabeled tree shapes compatible with $\pi = (2,2,3,3)$ is:

$$
\begin{aligned}
|\mathcal{G}_c^T(2,2,3,3)| &= |\mathcal{G}_c^T((2,2),3,3)| + |\mathcal{G}_c^T(2,2,(3,3))| + |\mathcal{G}_c^T((2,3),2,3)| - |\mathcal{G}_c^T((2,2),(3,3))| \\
&= \left[|\mathcal{G}_c^T((2,2),(3,3))| + |\mathcal{G}_c^T(((2,2),3),3)|\right] + \left[|\mathcal{G}_c^T((2,2),(3,3))| + |\mathcal{G}_c^T(((3,3),2),2)|\right] \\
&\quad + \left[|\mathcal{G}_c^T(((2,3),2),3)| + |\mathcal{G}_c^T(((2,3),3),2)| + |\mathcal{G}_c^T((2,3),(2,3))|\right] - |\mathcal{G}_c^T((2,2),(3,3))| \\
&= |\mathcal{G}_c^T((2,2),(3,3))| + |\mathcal{G}_c^T(((2,2),3),3)| + |\mathcal{G}_c^T(((3,3),2),2)| + |\mathcal{G}_c^T(((2,3),2),3)| \\
&\quad + |\mathcal{G}_c^T(((2,3),3),2)| + |\mathcal{G}_c^T((2,3),(2,3))| \\
&= 168 + 280 + 144 + 420 + 360 + 315 = 1687.
\end{aligned}
$$

In obtaining this sum, in intermediate steps, we use the fact that the values of $\mathcal{G}_c^T$ for (2), (3), (2,2), (3,3), (2,3), ((2,2),3), ((3,3),2), ((2,3),2)), ((2,3),3), and are 1, 1, 1, 3, 3, 10, 18, 15, and 45, respectively.

For counting the number of unranked unlabeled tree shapes compatible with $\pi = (n_1, n_2, \ldots, n_k)$, we simply replace $\mathcal{G}_c^T$ with $\mathcal{G}_c$ in eq. 18. We use Propositions 6 and 7 in place of Propositions 9 and 10.

Example 19. The number of unranked unlabeled tree shapes compatible with $\pi = (2, 2, 3, 3)$ is:

$$
\begin{aligned}
|\mathcal{G}_c(2, 2, 3, 3)| &= |\mathcal{G}_c((2, 2), (3, 3))| + |\mathcal{G}_c(((2, 2), 3), 3)| + |\mathcal{G}_c(((3, 3), 2), 2)| \\
&\quad + |\mathcal{G}_c(((2, 3), 2), 3)| + |\mathcal{G}_c(((2, 3), 3), 2)| + |\mathcal{G}_c((2, 3), (2, 3))| \\
&= 1 + 1 + 1 + 1 + 1 + 1 = 6.
\end{aligned}
$$

This example is quite straightforward; the values of $\mathcal{G}_c$ for the perfect phylogenies that appear in intermediate steps—(2), (3), (2,2), (3,3), (2,3), ((2,2),3), ((3,3),2), ((2,3),2)), and ((2,3),3)—all equal 1.

To count the number of ranked labeled tree shapes compatible with a labeled multifurcating perfect phylogeny $\pi = (n_1, n_2, \ldots, n_k)$, we assume that although any leaf in the perfect phylogeny can have multiplicity larger than one, each leaf is uniquely defined by its associated samples, all of which are all assumed to have different labels. Therefore, we take $\mathbf{a} = (n_1, n_2, \ldots, n_k)$ and $\mathbf{m} = (1, 1, \ldots, 1)$. Eq. 18 reduces to

$$
|\mathcal{G}_c^L(\pi)| = \underbrace{\sum_{i=1}^{s-1} \sum_{j=i+1}^{s} |\mathcal{G}_c^L(\mathcal{B}_{i,j}(\pi))|\, 1_{m_i > 0}\, 1_{m_j > 0}}_{\text{collapsing two pendant edges}}
$$

$$
- \underbrace{\sum_{i=1}^{s-1} \sum_{j=i+1}^{s} \sum_{\substack{k=1 \\ k \neq i, k \neq j}}^{s-1} \sum_{\substack{\ell=k+1 \\ \ell \neq i, \ell \neq j}}^{s} |\mathcal{G}_c^L(\mathcal{B}_{i,j}(\pi) \wedge \mathcal{B}_{k,\ell}(\pi))|\, 1_{m_i > 0}\, 1_{m_j > 0}\, 1_{m_k > 0}\, 1_{m_\ell > 0}}_{\text{collapsing two pairs of pendant edges}}. \qquad (19)
$$

The enumeration makes use of Proposition 13.

Example 20. Consider a labeled multifurcating perfect phylogeny that groups 2, 2, 3, and 3 samples at the root. We assume that $\mathbf{a} = (a_1, a_2, a_3, a_4) = (2, 2, 3, 3)$. Applying the recursion formula in eq. 19, we get

$$
\begin{aligned}
|\mathcal{G}_c^L(a_1, a_2, a_3, a_4)| &= |\mathcal{G}_c^L((a_1, a_2), a_3, a_4)| + |\mathcal{G}_c^L((a_1, a_3), a_2, a_4)| + |\mathcal{G}_c^L((a_1, a_4), a_2, a_3)| \\
&\quad + |\mathcal{G}_c^L((a_2, a_3), a_1, a_4)| + |\mathcal{G}_c^L((a_2, a_4), a_1, a_3)| + |\mathcal{G}_c^L((a_3, a_4), a_1, a_2)| \\
&\quad - |\mathcal{G}_c^L((a_1, a_2), (a_3, a_4))| - |\mathcal{G}_c^L((a_1, a_3), (a_2, a_4))| - |\mathcal{G}_c^L((a_1, a_4), (a_2, a_3))| \\
&= |\mathcal{G}_c^L((2, 2), 3, 3)| + 4|\mathcal{G}_c^L((2, 3), 2, 3)| + |\mathcal{G}_c^L((3, 3), 2, 2)| \\
&\quad - |\mathcal{G}_c^L((2, 2), (3, 3))| - 2|\mathcal{G}_c^L((2, 3), (2, 3))|.
\end{aligned}
$$

Now, because

$$
|\mathcal{G}_c^L(a_1, a_2, a_3)| = |\mathcal{G}_c^L((a_1, a_2), a_3)| + |\mathcal{G}_c^L((a_1, a_3), a_2)| + |\mathcal{G}_c^L((a_2, a_3), a_1)|,
$$

we have

$$
\begin{aligned}
|\mathcal{G}_c^L((2, 2), 3, 3)| &= 2|\mathcal{G}_c^L(((2, 2), 3), 3)| + |\mathcal{G}_c^L((2, 2), (3, 3))| \\
|\mathcal{G}_c^L((2, 3), 2, 3)| &= |\mathcal{G}_c^L(((2, 3), 2), 3)| + |\mathcal{G}_c^L(((2, 3), 3), 2)| + |\mathcal{G}_c^L((2, 3), (2, 3))| \\
|\mathcal{G}_c^L((3, 3), 2, 2)| &= 2|\mathcal{G}_c^L(((3, 3), 2), 2)| + |\mathcal{G}_c^L((2, 2), (3, 3))|.
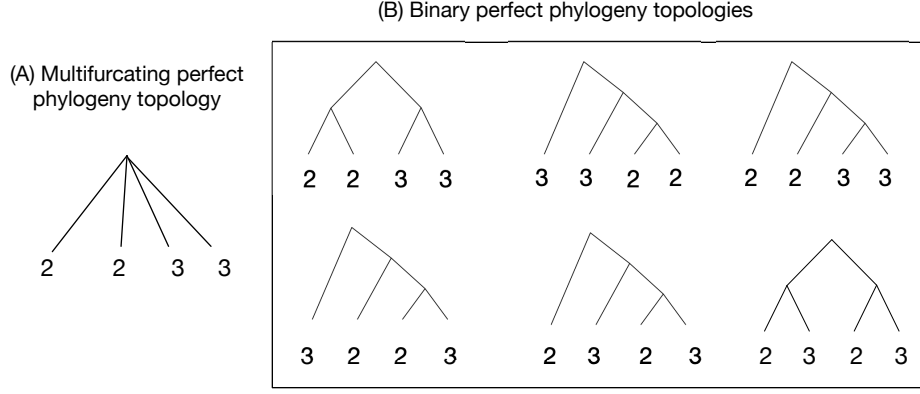\end{aligned}
$$

Figure 9: **Example of all possible binary perfect phylogeny topologies for a given multifurcating perfect phylogeny topology.** The binary perfect phylogenies are obtained from a multifurcating perfect phylogeny by resolving multifurcating nodes into sequences of bifurcations.

Summing all terms, we get

$$
\begin{aligned}
|\mathcal{G}_c^L(a_1, a_2, a_3, a_4)| &= 2|\mathcal{G}_c^L(((2,2),3),3)| + 2|\mathcal{G}_c^L(((3,3),2),2)| + |\mathcal{G}_c^L((2,2),(3,3))| \\
&+ 4|\mathcal{G}_c^L(((2,3),2),3)| + 4|\mathcal{G}_c^L(((2,3),3),2)| + 2|\mathcal{G}_c^L((2,3),(2,3))| \\
&= 2 \times 5040 + 2 \times 2592 + 6048 + 4 \times 3780 + 4 \times 3240 + 2 \times 5670 = 60,732.
\end{aligned}
$$

In obtaining this sum, we use the fact that the values of $\mathcal{G}_c^L$ for (2), (3), (2,2), (3,3), (2,3), ((2,2),3), ((3,3),2), ((2,3),2)), and ((2,3),3), and are 1, 3, 2, 54, 9, 60, 324, 45, and 405, respectively.

The number of unranked labeled tree shapes compatible with $\pi = (n_1, n_2, \ldots, n_k)$ is obtained by replacing $\mathcal{G}_c^L$ with $\mathcal{G}_c^X$ in eq. 19. We use Proposition 16 in place of Proposition 13.

Example 21. The number of unranked labeled tree shapes compatible with a labeled multifurcating perfect phylogeny that groups 2, 2, 3, and 3 samples at the root, with $\mathbf{a} = (a_1, a_2, a_3, a_4) = (2, 2, 3, 3)$ is:

$$
\begin{aligned}
|\mathcal{G}_c^X(a_1, a_2, a_3, a_4)| &= 2|\mathcal{G}_c^X(((2,2),3),3)| + 2|\mathcal{G}_c^X(((3,3),2),2)| + |\mathcal{G}_c^X((2,2),(3,3))| \\
&+ 4|\mathcal{G}_c^X(((2,3),2),3)| + 4|\mathcal{G}_c^X(((2,3),3),2)| + 2|\mathcal{G}_c^X((2,3),(2,3))| \\
&= 2 \times 9 + 2 \times 9 + 9 + 4 \times 9 + 4 \times 9 + 2 \times 9 = 135.
\end{aligned}
$$

The sum uses values of $\mathcal{G}_c^X$ for (2), (3), (2,2), (3,3), (2,3), ((2,2),3), ((3,3),2), ((2,3),2)), and ((2,3),3), equal to 1, 3, 1, 9, 3, 3, 9, 3, and 9, respectively.

5   Conclusion

The infinitely-many-sites mutations model is a popular model of molecular variation for problems of population genetics (Wakeley, 2008) and related areas (Jones et al., 2020), in which constraints are imposed on the space of trees that can explain the observed patterns of molecular variation. A realization of the coalescent model on a genealogy and a superimposed infinitely-many-sites mutation

Table 1: Number of trees compatible with example perfect phylogenies of 10 samples.

Perfect phylogeny	Unranked unlabeled tree shapes	Ranked unlabeled tree shapes	Ranked labeled tree shapes	Unranked labeled tree shapes
(9,1)	46	1385	57,153,600	2,027,025
(8,2)	23	2176	12,700,800	135,135
(7,3)	11	1708	4,762,800	31,185
(6,4)	12	1792	2,721,600	14,175
(5,5)	6	875	2,268,000	11,025
((8,1),1)	23	272	1,587,600	135,135
((7,2),1)	11	427	396,900	10,395
((6,3),1)	6	336	170,100	2835
((5,4),1)	6	350	113,400	1575
((7,1),2)	11	488	453,600	10,395
((6,2),2)	6	768	129,600	945
((5,3),2)	3	600	64,800	315
((4,4),2)	3	320	51,840	225
((6,1),3)	6	448	226,800	2835
((5,2),3)	3	700	75,600	315
((4,3),3)	2	560	45,360	135
((5,1),4)	6	560	181,440	1575
((4,2),4)	4	896	72,576	225
((3,3),4)	2	336	54,432	135
((4,1),5)	5	560	226,800	1575
((3,2),5)	3	735	113,400	315

The entries in the table are obtained by repeated use of Propositions 6 and 7 for unranked unlabeled tree shapes, 9 and 10 for ranked unlabeled tree shapes, 13 for ranked labeled tree shapes, and 16 for unranked labeled tree shapes. An arbitrary labeling of the perfect phylogeny is assumed for counting the associated ranked and unranked labeled tree shapes.

model can be summarized as a perfect phylogeny. Here, we have examined combinatorial properties of the genealogical tree structures that are compatible with a perfect phylogeny, demonstrating that the binary perfect phylogenies possess a lattice structure (Theorem 4). We have used this lattice structure to provide recursive enumerative results counting the trees—unranked unlabeled trees, ranked unlabeled trees, ranked labeled trees, and unranked labeled trees—compatible with binary and multifurcating perfect phylogenies.

In our enumerative results, the count of the number of trees of a specified type that are compatible with a perfect phylogeny is obtained by a decomposition of the perfect phylogeny at its root. The number of associated trees is obtained by counting trees for each subtree immediately descended from the root of the perfect phylogeny—and where appropriate, counting interleavings of nodes within those trees, taking care to consider cases that avoid double-counting, or both. This same technique was applicable for each of the types of trees we considered, appearing in Sections 3.2, 3.3, 3.4, 3.5, and 4. Owing to the recursive structure of the computation, the decomposition itself proceeds rapidly from the root through the internal nodes, so that a count can be quickly

obtained even if the number itself is large.

We obtained results concerning the cherry perfect phylogenies with the largest numbers of ranked unlabeled, unranked labeled, and ranked labeled tree shapes (Propositions 11, 14, and 17), and it will be informative to seek a similar result for the unranked unlabeled case. The result in Proposition 11 on asymptotic growth of the number of ranked unlabeled tree shapes compatible with a binary perfect phylogeny is reminiscent of a result concerning "lodgepole" trees. A number of studies have examined another combinatorial structure for evolutionary trees, the number of "coalescent histories" associated with a labeled species tree and its matching labeled gene tree. These coalescent histories encode different evolutionary scenarios possible for the coalescence of gene lineages on a species tree. Disanto and Rosenberg (2015) found that the lodgepole trees, a class of trees in which cherry nodes with 2 descendants successively branch from a single species tree edge, possesses a particularly large number of coalescent histories. Similarly, in Proposition 11, as n increases, the number of ranked unlabeled tree shapes compatible with a cherry perfect phylogeny is largest when the perfect phylogeny has one subtree with sample size 2.

Perfect phylogenies have been widely studied in varied estimation problems, for the "perfect phylogeny problem" asking whether a perfect phylogeny can be constructed from data given on a set of characters (Agarwala and Fernández-Baca, 1994; Kannan and Warnow, 1997; Felsenstein, 2004; Gusfield, 2014; Steel, 2016), statistical inference of evolutionary parameters under the coalescent (Griffiths and Tavaré, 1994; Stephens and Donnelly, 2000; Tavaré, 2004; Palacios et al., 2019; Cappello et al., 2020b), and algorithmic estimation of haplotype phase from diploid data (Gusfield, 2002; Bafna et al., 2004; Gusfield, 2014). However, the literature on perfect phylogenies has largely focused on such applications and on algorithmic problems of obtaining perfect phylogenies from data under various constraints, with little emphasis on the enumerative combinatorics of the perfect phylogenies themselves, and of their associated refinements. In describing a lattice for the binary perfect phylogenies with sample size n, this study suggests that the mathematical properties of sets of perfect phylogenies as combinatorial structures *per se* can be informative. The link to coalescent histories suggests possible connections to related concepts such as "ancestral configurations" (Wu, 2012; Disanto and Rosenberg, 2017), which also can be described in terms of lattices (E. Alimpiev & N.A.R., unpublished); it will be useful to consider perfect phylogenies alongside such structures arising in the combinatorics of evolutionary trees.

Finally, returning to considerations of coalescent-based inference from sequences, recall that inference of evolutionary parameters from a given perfect phylogeny is performed by integrating over the space of genealogies. A standard approach to inference integrates over the space of ranked labeled tree shapes generated by the Kingman coalescent (Drummond et al., 2012). However, this inference is computationally intractable for large sample sizes. We have observed a striking reduction in the cardinality of the set of ranked unlabeled tree shapes compatible with an observed perfect phylogeny, relative to the number of ranked labeled tree shapes compatible with an observed perfect phylogeny (Table 1). This observation contributes to a growing branch of the area of coalescent-based inference (Sainudiin et al., 2015; Palacios et al., 2015, 2019; Cappello et al., 2020a) that can make use of ranked unlabeled trees to estimate the evolutionary parameters.

6   Acknowledgments

References

R. Agarwala and D. Fernández-Baca. A faster algorithm for the perfect phylogeny problem when the number of characters is fixed. Computer Science Technical Reports, (5), 1994.

V. Bafna, D. Gusfield, S. Hannenhalli, and S. Yooseph. A note on efficient computation of haplotypes via perfect phylogeny. Journal of Computational Biology, 11(5):858–866, 2004.

R. P. Brent and D. Harvey. Fast computation of Bernoulli, Tangent and Secant numbers. In Computational and Analytical Mathematics, pages 127–142. Springer, 2013.

L. Cappello, J. A. Palacios, et al. Sequential importance sampling for multiresolution Kingman–Tajima coalescent counting. Annals of Applied Statistics, 14(2):727–751, 2020a.

L. Cappello, A. Veber, and J. A. Palacios. The Tajima heterochronous n-coalescent: inference from hete-rochronously sampled molecular data. arXiv preprint arXiv:2004.06826, 2020b.

F. Disanto and N. A. Rosenberg. Coalescent histories for lodgepole species trees. Journal of Computational Biology, 22(10):918–929, 2015.

F. Disanto and N. A. Rosenberg. Enumeration of ancestral configurations for matching gene trees and species trees. Journal of Computational Biology, 24(9):831–850, 2017.

A. Drummond, M. Suchard, D. Xie, and A. Rambaut. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution, 29:1969–1973, 2012.

J. Felsenstein. Inferring phylogenies, volume 2. Sinauer Associates Sunderland, MA, 2004.

P. Flajolet and R. Sedgewick. Analytic Combinatorics. Cambridge University press, 2009.

R. C. Griffiths and S. Tavaré. Sampling theory for neutral alleles in a varying environment. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 344(1310):403–410, June 1994.

D. Gusfield. Efficient algorithms for inferring evolutionary trees. Networks, 21(1):19–28, 1991.

D. Gusfield. Haplotyping as perfect phylogeny: conceptual framework and efficient solutions. In Proceedings of the sixth annual international conference on Computational biology, pages 166–175, 2002.

D. Gusfield. ReCombinatorics: the algorithmics of ancestral recombination graphs and explicit phylogenetic networks. MIT press, 2014.

E. F. Harding. The probabilities of rooted tree-shapes generated by random bifurcation. Advances in Applied Probability, 3(3):44–77, 1971.

S. Janson and G. Kersting. On the total external length of the Kingman coalescent. Electronic Journal of Probability, 16:2203–2218, 2011.

M. G. Jones, A. Khodaverdian, J. J. Quinn, M. M. Chan, J. A. Hussmann, R. Wang, C. Xu, J. S. Weissman, and N. Yosef. Inference of single-cell phylogenies from lineage tracing data using cassiopeia. Genome Biology, 21:1–27, 2020.

S. Kannan and T. Warnow. A fast algorithm for the computation and enumeration of perfect phylogenies. SIAM Journal on Computing, 26(6):1749–1763, 1997.

M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics, 61(4):893–903, 1969.

J. Kingman. The coalescent. <u>Stochastic Processes and their Applications</u>, 13(3):235–248, 1982.

P. Marjoram and S. Tavaré. Modern computational approaches for analysing molecular genetic variation data. <u>Nature Reviews Genetics</u>, 7(10):759–770, 2006.

J. B. Nation. Notes on lattice theory, 1998.

R. Otter. The number of trees. <u>Annals of Mathematics</u>, pages 583–599, 1948.

J. A. Palacios, J. Wakeley, and S. Ramachandran. Bayesian nonparametric inference of population size changes from sequential genealogies. <u>Genetics</u>, 201(1):281–304, 2015.

J. A. Palacios, A. Véber, L. Cappello, Z. Wang, J. Wakeley, and S. Ramachandran. Bayesian estimation of population size changes by sampling Tajima's trees. <u>Genetics</u>, 213(3):967–986, 2019.

R. Sainudiin, T. Stadler, and A. Véber. Finding the best resolution for the Kingman-Tajima coalescent: Theory and applications. <u>Journal of Mathematical Biology</u>, pages 1–41, 2015.

R. P. Stanley. <u>Enumerative Combinatorics</u>, volume 1. Cambridge Studies in Advanced Mathematics, New York, 2 edition.

M. Steel. <u>Phylogeny: discrete and random processes in evolution</u>. SIAM, 2016.

M. Stephens and P. Donnelly. Inference in molecular population genetics. <u>Journal of the Royal Statistical Society: Series B (Statistical Methodology)</u>, 62(4):605–635, 2000.

F. Tajima. Evolutionary relationship of DNA sequences in finite populations. <u>Genetics</u>, 105(2):437–460, 1983.

S. Tavaré. Part I: Ancestral inference in population genetics. In <u>Lectures on Probability Theory and Statistics</u>, volume 1837 of <u>Lecture Notes in Mathematics</u>, pages 1–188. Springer Verlag, New York, 2004.

J. Wakeley. <u>Coalescent Theory: An Introduction</u>. Roberts & Company Publishers, Greenwood Village, June 2008.

Y. Wu. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. <u>Evolution: International Journal of Organic Evolution</u>, 66(3):763–775, 2012.

Appendix: Proof of Theorem 4

To prove Theorem 4, we must verify four pairs of conditions concerning perfect phylogenies $\pi \in \Pi_n \cup \{\emptyset\}$. Note that any binary perfect phylogeny $\pi \in \Pi_n \cup \{\emptyset\}$ is equal to $\emptyset$, (n), or (π_1, π_2) for two non-empty binary perfect phylogenies $\pi_1 \in \Pi_{n_1}$ and $\pi_2 \in \Pi_{n_2}$, where $1 \leq n_1, n_2 < n$ and $n_1 + n_2 = n$. Hence, we must demonstrate the four pairs of conditions for perfect phylogeny pairs that include $\emptyset$, (n), or both, and for perfect phylogeny pairs that include neither $\emptyset$ nor (n).

Because perfect phylogenies can be decomposed into smaller perfect phylogenies, we proceed by induction on n, with a base case of $n = 1$. In the inductive step we assume that $(\Pi_k \cup \{\emptyset\}, \wedge, \vee)$ is a lattice for all k, $1 \leq k < n$. We then verify that it follows that $(\Pi_n \cup \{\emptyset\}, \wedge, \vee)$ is a lattice. We start with Condition 2, which is trivial.

Condition 2: $\pi \wedge \sigma = \sigma \wedge \pi$ **and** $\pi \vee \sigma = \sigma \vee \pi$

For all n, condition 2 of the definition of a lattice is trivially satisfied, as the operations $\wedge$ and $\vee$ are symmetric by definition. In subsequent derivations, we frequently apply Condition 2 without always noting its application.

The $n = 1$ case for Conditions 1, 3, and 4

Consider $n = 1$, for which Π_1 contains only the perfect phylogeny (1), and $\Pi_1 \cup \{\emptyset\}$ contains only (1) and $\emptyset$. For $\Pi_1 \cup \{\emptyset\}$, demonstrating Condition 1 of the requirements for a lattice requires that we show $(1) \wedge (1) = (1)$, $\emptyset \wedge \emptyset = \emptyset$, $(1) \vee (1) = (1)$, and $\emptyset \vee \emptyset = \emptyset$. These four relations are true by parts (3), (1), (4), and (2) of Defn. 1, respectively.

Demonstrating Condition 3 requires that we verify a pair of conditions for each of the eight choices of (x, y, z) for $x, y, z \in \Pi_1 \cup \{\emptyset\}$. Demonstrating Condition 4 requires that we verify a pair of conditions for each of the four choices of (x, y). The 16 verifications for Condition 3 and eight verifications for Condition 4 all quickly follow by Defn. 1 (1-4). Hence, $(\Pi_1 \cup \{\emptyset\}, \wedge, \vee)$ is a lattice.

Condition 1: $\pi \wedge \pi = \pi$ **and** $\pi \vee \pi = \pi$

First, we demonstrate the first part of the condition. We see $\emptyset \wedge \emptyset = \emptyset$ by Defn. 1 (1) and $(n) \wedge (n) = (n)$ by Defn. 1 (3).

Consider $\pi = (\pi_1, \pi_2)$ for $\pi_1 \in \Pi_{n_1}$ and $\pi_2 \in \Pi_{n_2}$, where $1 \leq n_1, n_2 < n$ and $n_1 + n_2 = n$.

$$\begin{aligned}
\pi \wedge \pi &= (\pi_1, \pi_2) \wedge (\pi_1, \pi_2) \\
&= (\pi_1 \wedge \pi_1, \pi_2 \wedge \pi_2) \vee (\pi_1 \wedge \pi_2, \pi_2 \wedge \pi_1) \text{ by Defn. 1 (8)} \\
&= (\pi_1, \pi_2) \vee (\pi_1 \wedge \pi_2, \pi_1 \wedge \pi_2) \text{ by the inductive hypothesis.}
\end{aligned}$$

If $n_1 \neq n_2$, then we apply Defn. 1 (5), the convention $(\pi, \emptyset) = \emptyset$, and Defn. 1 (2), and we obtain $\pi \wedge \pi = (\pi_1, \pi_2) \vee (\emptyset, \emptyset) = (\pi_1, \pi_2) \vee \emptyset = (\pi_1, \pi_2) = \pi$. If $n_1 = n_2$, then we have two cases: $\pi_1 \leq \pi_2$ (without loss of generality), and π_1, π_2 are not comparable.

If $\pi_1 \leq \pi_2$, then $\pi_1 \wedge \pi_2 = \pi_1$ and $\pi_1 \vee \pi_2 = \pi_2$. By Defn. 1 (9), $(\pi_1, \pi_2) \vee (\pi_1, \pi_1) = (\pi_1, \pi_2 \vee \pi_1) = (\pi_1, \pi_2) = \pi$, so that $\pi \wedge \pi = \pi$.

If π_1 and π_2 are not comparable, then by Defn. 1 (11), $\pi_1 \wedge \pi_2 = \delta$ for some $\delta \in (\Pi_{n1} \cup \{\emptyset\}) \setminus \{\pi_1, \pi_2\}$, with $\delta \vee \pi_1 = \pi_1$ and $\delta \vee \pi_2 = \pi_2$. We then have by Defn. 1 (9),

$$(\pi_1, \pi_2) \vee (\pi_1 \wedge \pi_2, \pi_1 \wedge \pi_2) = (\pi_1, \pi_2) \vee (\delta, \delta).$$

But (δ, δ) refines (π_1, π_2), as δ refines π_1 and δ refines π_2, so that (π_1, π_2) can be obtained by collapsing cherries separately in the two subtrees of (δ, δ). Hence, $\pi \wedge \pi = (\pi_1, \pi_2) \vee (\delta, \delta) = (\pi_1, \pi_2) = \pi$.

For the second part of the condition, we have $\emptyset \vee \emptyset = \emptyset$ by Defn. 1 (2) and $(n) \vee (n) = (n)$ by Defn. 1 (4). Consider $\pi = (\pi_1, \pi_2)$ for $\pi_1 \in \Pi_{n_1}$ and $\pi_2 \in \Pi_{n_2}$, where $1 \leq n_1, n_2 < n$ and $n_1 + n_2 = n$.

$$\begin{aligned}
\pi \vee \pi &= (\pi_1, \pi_2) \vee (\pi_1, \pi_2) \\
&= (\pi_1, \pi_2 \vee \pi_2) \text{ by Defn. 1 (9)} \\
&= (\pi_1, \pi_2) \text{ by the inductive hypothesis} \\
&= \pi.
\end{aligned}$$

Condition 4: $\pi \wedge (\pi \vee \sigma) = \pi$ **and** $\pi \vee (\pi \wedge \sigma) = \pi$

First, we see that both parts of the condition hold if at least one of π, σ is in $\{\emptyset, (n)\}$, by Defn. 1 (1-4). Next, we have the following 3 cases:

i. If $\pi \leq \sigma$, then $\pi \wedge \sigma = \pi$ and $\pi \vee \sigma = \sigma$. Hence, $\pi \wedge (\pi \vee \sigma) = \pi \wedge \sigma = \pi$. By Condition 1, $\pi \vee (\pi \wedge \sigma) = \pi \vee \pi = \pi$.

ii. If $\sigma \leq \pi$, then $\pi \wedge \sigma = \sigma$ and $\pi \vee \sigma = \pi$. Hence, by Condition 1, $\pi \wedge (\pi \vee \sigma) = \pi \wedge \pi = \pi$. We also have $\pi \vee (\pi \wedge \sigma) = \pi \vee \sigma = \pi$.

iii. If π and σ are not comparable, then by Defn. 1 (11), there exists a perfect phylogeny γ such that $\pi \vee \sigma = \gamma$, $\pi \wedge \gamma = \pi$, and $\sigma \wedge \gamma = \sigma$. Hence $\pi \wedge (\pi \vee \sigma) = \pi \wedge \gamma = \pi$. By Defn. 1 (11), there exists a perfect phylogeny ρ such that $\pi \wedge \sigma = \rho$, $\pi \vee \rho = \pi$, and $\sigma \vee \rho = \sigma$. We have $\pi \vee (\pi \wedge \sigma) = \pi \vee \rho = \pi$.

Condition 3: $\pi \wedge (\sigma \wedge \rho) = (\pi \wedge \sigma) \wedge \rho$ **and** $\pi \vee (\sigma \vee \rho) = (\pi \vee \sigma) \vee \rho$

First, we see that both parts of the condition hold if at least one of π, σ, ρ is in $\{\emptyset, (n)\}$, by Defn. 1 (1-4). Assume now that $\pi = (\pi_1, \pi_2)$, $\sigma = (\sigma_1, \sigma_2)$, and $\rho = (\rho_1, \rho_2)$. Then

$$
\begin{aligned}
\pi \wedge (\sigma \wedge \rho) &= (\pi_1, \pi_2) \wedge ((\sigma_1, \sigma_2) \wedge (\rho_1, \rho_2)) \\
&= (\pi_1, \pi_2) \wedge [(\sigma_1 \wedge \rho_1, \sigma_2 \wedge \rho_2) \vee (\sigma_1 \wedge \rho_2, \sigma_2 \wedge \rho_1)] \text{ by Defn. 1 (8)} \\
&= [(\pi_1, \pi_2) \wedge (\sigma_1 \wedge \rho_1, \sigma_2 \wedge \rho_2)] \vee [(\pi_1, \pi_2) \wedge (\sigma_1 \wedge \rho_2, \sigma_2 \wedge \rho_1)] \text{ by Defn. 1 (10)} \\
&= [(\pi_1 \wedge (\sigma_1 \wedge \rho_1), \pi_2 \wedge (\sigma_2 \wedge \rho_2)) \vee (\pi_1 \wedge (\sigma_2 \wedge \rho_2), \pi_2 \wedge (\sigma_1 \wedge \rho_1))] \\
&\quad \vee [(\pi_1 \wedge (\sigma_1 \wedge \rho_2), \pi_2 \wedge (\sigma_2 \wedge \rho_1)) \vee (\pi_1 \wedge (\sigma_2 \wedge \rho_1), \pi_2 \wedge (\sigma_1 \wedge \rho_2))] \text{ by Defn. 1 (8)}
\end{aligned}
$$

By the inductive hypothesis *for both parts of the condition*, $\pi_i \wedge (\sigma_j \wedge \rho_k) = (\pi_i \wedge \sigma_j) \wedge \rho_k$ and $\pi_i \vee (\sigma_j \vee \rho_k) = (\pi_i \vee \sigma_j) \vee \rho_k$ for all $i, j, k \in \{1, 2\}$. We then get

$$
\begin{aligned}
\pi \wedge (\sigma \wedge \rho) &= [((\pi_1 \wedge \sigma_1) \wedge \rho_1, (\pi_2 \wedge \sigma_2) \wedge \rho_2) \vee ((\pi_1 \wedge \sigma_2) \wedge \rho_2, (\pi_2 \wedge \sigma_1) \wedge \rho_1)] \\
&\quad \vee [((\pi_1 \wedge \sigma_1) \wedge \rho_2, (\pi_2 \wedge \sigma_2) \wedge \rho_1) \vee ((\pi_1 \wedge \sigma_2) \wedge \rho_1, (\pi_2 \wedge \sigma_1) \wedge \rho_2)].
\end{aligned}
$$

By the inductive hypothesis for operator $\vee$ and by Condition 2, we can rearrange parentheses and swap the order of terms to obtain:

$$
\begin{aligned}
\pi \wedge (\sigma \wedge \rho) &= ((\pi_1 \wedge \sigma_1) \wedge \rho_1, (\pi_2 \wedge \sigma_2) \wedge \rho_2) \vee [(\pi_1 \wedge \sigma_1) \wedge \rho_2, (\pi_2 \wedge \sigma_2) \wedge \rho_1) \\
&\quad \vee ((\pi_1 \wedge \sigma_2) \wedge \rho_2, (\pi_2 \wedge \sigma_1) \wedge \rho_1] \vee ((\pi_1 \wedge \sigma_2) \wedge \rho_1, (\pi_2 \wedge \sigma_1) \wedge \rho_2).
\end{aligned}
$$

Dropping the brackets and viewing this expression as having four perfect phylogenies separated by the $\vee$ operator, we group the first two and the last two perfect phylogenies together and apply Defn. 1 (8) to each group. We get

$$
\begin{aligned}
\pi \wedge (\sigma \wedge \rho) &= [(\pi_1 \wedge \sigma_1, \pi_2 \wedge \sigma_2) \wedge (\rho_1, \rho_2)] \vee [(\pi_1 \wedge \sigma_2, \pi_2 \wedge \sigma_1) \wedge (\rho_1, \rho_2)] \text{ by Defn. 1 (8)} \\
&= [(\pi_1 \wedge \sigma_1, \pi_2 \wedge \sigma_2) \vee (\pi_1 \wedge \sigma_2, \pi_2 \wedge \sigma_1)] \wedge (\rho_1, \rho_2) \text{ by Defn. 1 (10)} \\
&= (\pi \wedge \sigma) \wedge \rho \text{ by Defn. 1 (8).}
\end{aligned}
$$

For the second part of the condition, suppose $\pi = (\pi_1, \pi_2) \in \Pi_n$, $\sigma = (\sigma_1, \sigma_2) \in \Pi_n$, and $\rho = (\rho_1, \rho_2) \in \Pi_n$ are three perfect phylogenies of size n. We consider four cases. First, suppose the three perfect phylogenies have mutually different subtree sizes—that is, $\{|\pi_1|, |\pi_2|\}$, $\{|\sigma_1|, |\sigma_2|\}$, and $\{|\rho_1|, |\rho_2|\}$ are mutually distinct. Then $\pi \vee \sigma = \pi \vee \rho = \sigma \vee \rho = (n)$ by Defn. 1 (9). We then have $\pi \vee (\sigma \vee \rho) = \pi \vee (n) = (n) = (n) \vee \rho = (\pi \vee \sigma) \vee \rho$ by Defn. 1 (4).

The same argument applies if it is merely assumed that σ and ρ have pairs of subtrees whose sizes differ, $\{|\sigma_1|, |\sigma_2|\} \neq \{|\rho_1|, |\rho_2|\}$). Then $\pi \vee (\sigma \vee \rho) = \pi \vee (n) = (n) = (\pi \vee \sigma) \vee \rho$ by Defn. 1 (4, 9), where we have used the fact that $\sigma \vee \rho = (n)$ and $\sigma \leq \pi \vee \sigma$, so that $(\pi \vee \sigma) \vee \rho = (n)$.

If $\{|\sigma_1|, |\sigma_2|\} = \{|\rho_1|, |\rho_2|\}$) but $\{|\pi_1|, |\pi_2|\} \neq \{|\sigma_1|, |\sigma_2|\}$) and $\{|\pi_1|, |\pi_2|\} \neq \{|\rho_1|, |\rho_2|\}$), then $\pi \vee \sigma = (n)$. Because $\sigma \leq \sigma \vee \rho$ and $\pi \vee \sigma = (n)$, $\pi \vee (\sigma \vee \rho) = (n)$. Similarly, $(\pi \vee \sigma) \vee \rho = (n) \vee \rho = (n)$ by Defn. 1 (4, 9).

It remains to consider the case in which at least a pair of subtrees, one each from π, σ and ρ have the same size, or $\{|\pi_1|, |\pi_2|\} = \{|\sigma_1|, |\sigma_2|\}) = \{|\rho_1|, |\rho_2|\}$). We have

$$
\begin{aligned}
\pi \vee (\sigma \vee \rho) &= (\pi_1, \pi_2) \vee ((\sigma_1, \sigma_2) \vee (\rho_1, \rho_2)) \\
&= (\pi_1, \pi_2) \vee [(\sigma_1 \vee \rho_1, \sigma_2 \vee \rho_2) \wedge (\sigma_1 \vee \rho_2, \sigma_2 \vee \rho_1)] \text{ by Defn. 1 (9)} \\
&= [(\pi_1, \pi_2) \vee (\sigma_1 \vee \rho_1, \sigma_2 \vee \rho_2)] \wedge [(\pi_1, \pi_2) \vee (\sigma_1 \vee \rho_2, \sigma_2 \vee \rho_1)] \text{ by Defn. 1 (10)} \\
&= [(\pi_1 \vee (\sigma_1 \vee \rho_1), \pi_2 \vee (\sigma_2 \vee \rho_2)) \wedge (\pi_1 \vee (\sigma_2 \vee \rho_2), \pi_2 \vee (\sigma_1 \vee \rho_1))] \\
&\quad \wedge [(\pi_1 \vee (\sigma_1 \vee \rho_2), \pi_2 \vee (\sigma_2 \vee \rho_1)) \wedge (\pi_1 \vee (\sigma_2 \vee \rho_1), \pi_2 \vee (\sigma_1 \vee \rho_2))] \text{ by Defn. 1 (9)} \\
&= ((\pi_1 \vee \sigma_1) \vee \rho_1, (\pi_2 \vee \sigma_2) \vee \rho_2) \wedge ((\pi_1 \vee \sigma_1) \vee \rho_2, (\pi_2 \vee \sigma_2) \vee \rho_1) \\
&\quad \wedge ((\pi_1 \vee \sigma_2) \vee \rho_2, (\pi_2 \vee \sigma_1) \vee \rho_1) \wedge ((\pi_1 \vee \sigma_2) \vee \rho_1, (\pi_2 \vee \sigma_1) \vee \rho_2) \text{ by ind. hypothesis} \\
&= [(\pi_1 \vee \sigma_1, \pi_2 \vee \sigma_2) \vee (\rho_1, \rho_2)] \wedge [(\pi_1 \vee \sigma_2, \pi_2 \vee \sigma_1) \vee (\rho_1, \rho_2)] \text{ by Defn. 1 (9)} \\
&= [(\pi_1 \vee \sigma_1, \pi_2 \vee \sigma_2) \wedge (\pi_1 \vee \sigma_2, \pi_2 \vee \sigma_1)] \vee (\rho_1, \rho_2) \text{ by Defn. 1 (10)} \\
&= (\pi \vee \sigma) \vee \rho \text{ by Defn. 1 (9).} \quad\quad (20)
\end{aligned}
$$

Note that this derivation includes the case of shared subtrees at the root, in which it is not only the sizes of the subtrees that are the same, but the subtrees themselves. For example, suppose $\pi = (\pi_1, \pi_2)$ and $\sigma = (\pi_1, \sigma_1)$. By Defn. 1 (9), we have

$$
\pi \vee \sigma = (\pi_1, \pi_2) \vee (\pi_1, \sigma_1) = (\pi_1, \pi_2 \vee \sigma_1).
$$

However, we will show that we can replace the previous equality by the extended expression:

$$
\pi \vee \sigma = (\pi_1 \vee \pi_1, \pi_2 \vee \sigma_1) \wedge (\pi_1 \vee \sigma_1, \pi_1 \vee \pi_2), \quad\quad (21)
$$

and then the previous derivation remains unchanged. To prove this assertion, we have:

$$
\begin{aligned}
\pi_2 \vee \sigma_1 &= (\pi_2 \vee (\pi_1 \wedge \pi_2)) \vee \sigma_1 \quad \text{by Condition 4} \\
&= (\pi_2 \vee \sigma_1) \vee (\pi_1 \wedge \pi_2) \quad \text{by the inductive hypothesis and Condition 2} \\
&= \pi_2 \vee [(\sigma_1 \wedge \pi_1) \vee \sigma_1] \vee (\pi_1 \wedge \pi_2) \quad \text{by Conditions 2 and 4} \\
&= [\pi_2 \vee (\sigma_1 \wedge \pi_1)] \vee [\sigma_1 \vee (\pi_1 \wedge \pi_2)] \quad \text{by the inductive hypothesis.}
\end{aligned}
$$

Then

$$\pi \vee \sigma = (\pi_1, \pi_2 \vee \sigma_1)$$
$$= (\pi_1, [\pi_2 \vee (\sigma_1 \wedge \pi_1)] \vee [\sigma_1 \vee (\pi_1 \wedge \pi_2)])$$
$$= (\pi_1, \pi_2 \vee (\sigma_1 \wedge \pi_1)) \vee (\pi_1, \sigma_1 \vee (\pi_1 \wedge \pi_2)) \quad \text{by Defn. 1 (9)}$$
$$= (\pi_1, (\pi_2 \vee \sigma_1) \wedge (\pi_2 \vee \pi_1)) \vee (\pi_1, (\sigma_1 \vee \pi_1) \wedge (\sigma_1 \vee \pi_2)) \quad \text{by Defn. 1 (10).} \qquad (22)$$

By Condition 4 and Defn. 1 (9) we have

$$\pi_1 = \pi_1 \vee (\pi_1 \wedge \sigma_1) = (\pi_1 \vee \pi_1) \wedge (\pi_1 \vee \sigma_1),$$

and

$$\pi_1 = \pi_1 \vee (\pi_1 \wedge \pi_2) = (\pi_1 \vee \pi_1) \wedge (\pi_1 \vee \pi_2).$$

Replacing the first π_1 in the first pair of eq. 22 by $(\pi_1 \vee \pi_1) \wedge (\pi_1 \vee \sigma_1)$, and the first π_1 in the second pair of eq. 22 by $(\pi_1 \vee \pi_1) \wedge (\pi_1 \vee \pi_2)$, we get

$$\pi \vee \sigma = ((\pi_1 \vee \pi_1) \wedge (\pi_1 \vee \sigma_1), (\pi_2 \vee \sigma_1) \wedge (\pi_2 \vee \pi_1)) \vee ((\pi_1 \vee \pi_1) \wedge (\pi_1 \vee \pi_2), (\sigma_1 \vee \pi_1) \wedge (\sigma_1 \vee \pi_2))$$
$$= (\pi_1 \vee \pi_1, \pi_2 \vee \sigma_1) \wedge (\pi_1 \vee \sigma_1, \pi_1 \vee \pi_2) \quad \text{by Defn. 1 (8).}$$

Thus, eq. 21 holds, so that eq. 20 holds for the case in which subtrees are shared at the root.