Enumeration of binary trees compatible with a perfect phylogeny

Julia A. Palacios 1,2, Anand Bhaskar 3, Filippo Disanto 4, Noah A. Rosenberg 5

Received: 10 August 2021 / Revised: 7 March 2022 / Accepted: 31 March 2022 / Published online: 12 May 2022
© The Author(s) 2022

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.

Julia A. Palacios
juliapr@stanford.edu

1 Department of Statistics, Stanford University, Stanford, CA, USA
2 Department of Biomedical Data Science, Stanford Medicine, Stanford, CA, USA
3 Department of Genetics, Stanford University, Stanford, CA, USA
4 Department of Mathematics, University of Pisa, Pisa, Italy
5 Department of Biology, Stanford University, Stanford, CA, USA
Mathematics Subject Classification 05C05, 92D15

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 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/1cAx2xyn70tmG-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 \( \frac{k}{2} \) 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 \( i \)th 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 as a ranked unlabeled tree shape with internal nodes labeled by the transition index.

A ranked labeled tree shape of size \( n \), denoted by \( T^L_n \), 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 \( T^L_n \). 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 \( T^R_n \), 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 $T_n^R$. 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 $T_n^U$, 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 $T_n^U$. 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 $T_n^L$, 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 $T_n^L$. 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 mutation 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 sequence, with 0 denoting the ancestral type and 1 denoting the mutant type at any site; we assume that the ancestral type is known, and that it is denoted by 0.

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 rooted 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 Fig. 3A and B 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.

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 Fig. 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 \( \Pi_n \) 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 \( \pi \) 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 \( \pi_1 \) on the right in Fig. 3B in parenthetical notation can be written \((2, (2, 1))\) or \(((2, 1), 2)\), 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 \( \pi \) if the tree can be reduced to \( \pi \) 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 \( L_n = |T^L_n| \) 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, 56,700.

Let \( X_n = |T^X_n| \) 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 \( T^X_n \) from trees in \( T^X_{n-1} \), a pendant edge connected to the \( n \)th 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, 10,395.

The number of ranked tree shapes with \( n \) tips is the \((n-1)\)-th Euler zigzag number (Stanley 2012). Let \( R_n = |T^R_n| \) denote the number of ranked tree shapes with \( n \) leaves. We have the following recursion:

\[
R_1 = 1, \quad R_2 = 1,
\]

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

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 \( S_n = |T^L_n| \) denote the number of unranked tree shapes with \( n \) leaves. We have the following recursion:

\[
S_1 = 1,
\]
\[ S_{2n-1} = \sum_{k=1}^{n-1} S_k S_{2n-1-k}, \quad n \geq 2, \]  
\[ S_{2n} = \left( \sum_{k=1}^{n-1} S_k S_{2n-k} \right) + \frac{1}{2} S_n (S_n + 1), \quad n \geq 1. \]

\( S_n \) is the \( n \)th 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 Sect. 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 the set \( \Pi_n \) of all binary perfect phylogenies of \([n]\) into a poset by defining \( \pi \leq \sigma \) if either \( \sigma \) is the same as \( \pi \), or if \( \sigma \) can be obtained by sequentially collapsing pairs of pendant edges, or cherries, of \( \pi \). We then say \( \pi \) is a refinement of \( \sigma \). For example, \( \pi = (2, 3) \) refines \( \sigma = (5) \). We say that two binary perfect phylogenies in \( \Pi_n \) 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 \( \pi_1 \) and \( \pi_2 \) in \( \Pi_n \), their meet, denoted \( \pi_1 \land \pi_2 \), is the largest perfect phylogeny that refines both \( \pi_1 \) and \( \pi_2 \). Similarly, the join of two binary perfect phylogenies \( \pi_1 \lor \pi_2 \) is the smallest perfect phylogeny that is refined by both \( \pi_1 \) and \( \pi_2 \). Formal definitions of these notions appear in Definition 1.

Under the meet and join operations, we will see in Theorem 5 that the poset \( \Pi_n \cup \{\emptyset\} \) is a lattice \( \Lambda_n = (\Pi_n \cup \{\emptyset\}, \land, \lor) \). As a lattice, \( \Lambda_n \) possesses a Hasse diagram with a minimal and a maximal element. The maximal element of \( \Lambda_n \) is the single node perfect phylogeny \( (n) \) and the minimal element is \( \emptyset \). Figures 4 and 5 show the Hasse diagrams of \( \Lambda_2, \Lambda_3, \Lambda_4, \Lambda_5 \).

**Definition 1** (Binary perfect phylogeny operations). We define the binary perfect phylogeny symmetric operations \( \land, \lor : (\bigcup_{n \geq 1} \Pi_n \cup \{\emptyset\}) \times (\bigcup_{n \geq 1} \Pi_n \cup \{\emptyset\}) \to \Lambda_n \).
Fig. 4 Hasse diagrams of the lattices of binary perfect phylogenies with $n = 2, 3,$ and $4$ taxa

Fig. 5 Hasse diagram of the lattice of binary perfect phylogenies with $n = 5$ taxa

$(\cup_{n \geq 1} \Pi_n \cup \{\emptyset\})$, where $\Pi_n$ is the space of binary perfect phylogenies of $n$ leaves, as follows:

1. $\pi \land \emptyset = \emptyset$, for all $\pi \in \Pi_n$.
2. $\pi \lor \emptyset = \pi$, for all $\pi \in \Pi_n$.
3. $\pi \land (n) = \pi$, for all $\pi \in \Pi_n$.
4. $\pi \lor (n) = (n)$, for all $\pi \in \Pi_n$.
5. $\pi_1 \land \pi_2 = \emptyset$, for all $\pi_1 \in \Pi_{n_1}, \pi_2 \in \Pi_{n_2}$, with $n_1 \neq n_2$.
6. $\pi_1 \lor \pi_2 = \emptyset$, for all $\pi_1 \in \Pi_{n_1}, \pi_2 \in \Pi_{n_2}$, with $n_1 \neq n_2$.

The following proposition extends properties of the meet and join operations. It is proved in the “Appendix”.
Proposition 2 Let $\wedge, \lor$ be the two binary perfect phylogeny operations of Definition 1. Then:

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

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

2. For all $\pi_1, \pi_2, \pi_3, \pi_4$ with $(\pi_1, \pi_2) \in \Pi_n$ and $(\pi_3, \pi_4) \in \Pi_n$,

$$(\pi_1, \pi_2) \land (\pi_3, \pi_4) = (\pi_1 \land \pi_3, \pi_2 \land \pi_4) \lor (\pi_1 \land \pi_4, \pi_2 \land \pi_3),$$

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.

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

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

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 (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, 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 $(n)$.

4. For all $\pi_1, \pi_2, \pi_3 \in \Pi_n$,

$$\pi_1 \land (\pi_2 \lor \pi_3) = (\pi_1 \land \pi_2) \lor (\pi_1 \land \pi_3),$$

and

$$\pi_1 \lor (\pi_2 \land \pi_3) = (\pi_1 \lor \pi_2) \land (\pi_1 \lor \pi_3).$$
5. 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

\[ \pi \land \sigma = \gamma, \quad \pi \lor \gamma = \pi, \quad \text{and} \quad \sigma \lor \gamma = \sigma, \]

and

\[ \pi \lor \sigma = \rho, \quad \pi \land \rho = \pi, \quad \text{and} \quad \sigma \land \rho = \sigma. \]

Note that the meet and join operations are symmetric and that pairs \( (\pi_1, \pi_2) \) are unordered; for convenience, we have expanded expressions in parts 1 and 3 of the proposition that could potentially be simplified using the symmetry.

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

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

\[
((4, 2), 6) \land ((3, 3), 6) \\
= (4, 2) \land (3, 3, 6 \land 6) \lor ((4, 2) \land 6, 6 \land (3, 3)) \quad \text{by Prop. 2 (2)} \\
= (\emptyset, 6) \lor ((4, 2), (3, 3)) \quad \text{by Defn. 1 (3, 5) and Prop. 2 (2)} \\
= 0 \lor ((4, 2), (3, 3)) \quad \text{by convention} \\
= ((4, 2), (3, 3)) \quad \text{by Defn. 1 (2)}.
\]

\[
((4, 2), 6) \lor ((3, 3), 6) \\
= (6, (4, 2) \lor (3, 3)) \quad \text{by Prop. 2 (3)} \\
= (6, 6) \quad \text{by Prop. 2 (1)}.
\]

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

\[
(((3, 1), 2, 6) \land ((4, 2), 6) \\
= (((3, 1), 2) \land (4, 2), 6 \land 6) \lor (((3, 1), 2) \land 6, 6 \land (4, 2)) \quad \text{by Prop. 2 (2)} \\
= (((3, 1), 2) \land (4, 2), 6) \lor (((3, 1), 2), (4, 2)) \quad \text{by Defn. 1 (3)} \\
= (((3, 1) \land 4, 2 \land 2), 6) \lor (((3, 1), 2), (4, 2)) \quad \text{by Defn. 1 (2, 5) and Prop. 2 (2)} \\
= (((3, 1), 2), 6) \lor (((3, 1), 2), (4, 2)) \quad \text{by Defn. 1 (3)} \\
= (((3, 1), 2), 6) \quad \text{by Defn. 1 (4) and Prop. 2 (3)}.
\]

\[
((3, 1), 2, 6) \lor ((4, 2), 6) \\
= (((3, 1), 2) \lor (4, 2), 6) \quad \text{by Prop. 2 (3)} \\
= ((4, 2), 6) \quad \text{by Defn. 1 (4) and Prop. 2 (3)}.
\]

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

A lattice (Nation 1998) is an algebra \( \mathcal{L}(L, \land, \lor) \) satisfying, for all \( x, y, z \in L \),
Fig. 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 \(((3, 1), 6)\) and \(((4, 2), 6)\), their meet is \(((3, 1), (2), 6)\) and their join is \(((4, 2), 6)\).

1. \(x \land x = x\) and \(x \lor x = x\),
2. \(x \land y = y \land x\) and \(x \lor y = y \lor x\),
3. \(x \land (y \land z) = (x \land y) \land z\) and \(x \lor (y \lor z) = (x \lor y) \lor z\),
4. \(x \land (x \lor y) = x\) and \(x \lor (x \land y) = x\).

In the “Appendix,” we verify these conditions for \((\Pi_n \cup \{\emptyset\}, \land, \lor)\), giving the following theorem.

**Theorem 5** \((\Pi_n \cup \{\emptyset\}, \land, \lor)\) 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 6** (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 \(\pi\) 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 \(\pi\).

We use the symbol \(\mathcal{G}_c(\pi) = \{T_n : T_n \sim \pi\}\) to denote the set of unranked unlabeled tree shapes compatible with a perfect phylogeny \(\pi \in \Pi_n\). For a perfect phylogeny \(\pi\) consisting of a single leaf with leaf count \(n\), the number of compatible unranked unlabeled tree shapes is simply the number of unranked unlabeled tree 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 7 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

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

Proof By Definition 6, 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 \(n_1 = n_2\).

Proposition 8 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

\[
|\mathcal{G}_c((\pi_1, \pi_2))| = \begin{cases} 
|\mathcal{G}_c(\pi_1)| \cdot |\mathcal{G}_c(\pi_2)| - \frac{1}{2}|\mathcal{G}_c(\pi_1 \land \pi_2)| \cdot |\mathcal{G}_c(\pi_1 \land \pi_2)| - 1 & \text{if } \pi_1 \land \pi_2 \neq \emptyset \\
|\mathcal{G}_c(\pi_1)| \cdot |\mathcal{G}_c(\pi_2)| & \text{if } \pi_1 \land \pi_2 = \emptyset.
\end{cases}
\] (7)

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

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

\[\square\]
Propositions 7 and 8 provide a recursive formula for calculating the number of tree shapes compatible with a binary perfect phylogeny. For example, examining Fig. 6A, the number of tree shapes compatible with \((4, 2)\) is \(S_4S_2 = 2\), and the number of tree shapes compatible with \(((4, 2), 6)\) is \(|G_c(4, 2)| |G_c(6)| - \frac{1}{2} |G_c(4, 2)| (|G_c(4, 2)| - 1) = (2)(6) - \frac{1}{2}(2)(1) = 11\). 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 9** (Ranked unlabeled tree shape \(T^R_n\) compatible with a perfect phylogeny \(\pi \in \Pi_n\)). A ranked unlabeled tree shape with \(n\) leaves, \(T^R_n\), is compatible with a perfect phylogeny \(\pi \in \Pi_n\) if the unranked unlabeled tree shape \(T_n\) obtained by removing the ranking from \(T^R_n\) is compatible with \(\pi\).

**Proposition 10** 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

\[
|G^T_c((n_1, n_2))| = \left\{ \begin{array}{ll}
\binom{n_1+n_2-2}{n_1-1} R_{n_1} R_{n_2} & \text{if } n_1 \neq n_2 \\
\frac{1}{2} \binom{2n_1-2}{n_1-1} R^2_{n_1} & \text{if } n_1 = n_2.
\end{array} \right.
\] (8)

**Proof** By Definition 9, a ranked unlabeled tree shape \(T^R_n\) 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 \(\pi\). By Definition 6, 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 \(\frac{1}{2}(R^2_{n_1} - R_{n_1})\) ways, each with \(\binom{2n_1-2}{n_1-1}\) ways of interleaving the subtrees.

**Proposition 11** 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

\[
|G^T_c((\pi_1, \pi_2))| = \left\{ \begin{array}{ll}
\binom{2n_1-2}{n_1-1} \left( |G^T_c(\pi_1)| |G^T_c(\pi_2)| - \frac{1}{2} |G^T_c(\pi_1 \land \pi_2)|^2 \right) & \text{if } \pi_1 \land \pi_2 \neq \emptyset \\
\binom{n_1+n_2-2}{n_1-1} |G^T_c(\pi_1)| \left( |G^T_c(\pi_2)| \right) & \text{if } \pi_1 \land \pi_2 = \emptyset.
\end{array} \right.
\] (9)
Proposition 12

For a fixed integer value $\pi_1 \land \pi_2 = \emptyset$, then the number of ranked tree shapes compatible with $(\pi_1, \pi_2)$ is simply the product of the number of ranked tree shapes compatible with $\pi_1$, the number of ranked tree shapes compatible with $\pi_2$, and the number of ways of interleaving their rankings.

If $\pi_1 \land \pi_2 \neq \emptyset$, then certain ranked tree shapes can be compatible with both $\pi_1$ and $\pi_2$, i.e., compatible with $\pi_1 \land \pi_2$. 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 $\frac{1}{2} \left| G(\pi_1 \land \pi_2) \right|^2$, $|G(\pi_1) - |G(\pi_1 \land \pi_2)| + |G(\pi_1 \land \pi_2)||G(\pi_1)| - |G(\pi_1 \land \pi_2)|)$, and $|G(\pi_1)| - |G(\pi_1 \land \pi_2)|))|G(\pi_2)| - |G(\pi_1 \land \pi_2)|)$, respectively, all multiplied by the possible number of interleavings of the rankings $2^{n_1n_2}$.

Propositions 10 and 11 provide a recursive formula for calculating the number of ranked tree shapes compatible with a binary perfect phylogeny. For Fig. 6A, the number of ranked tree shapes compatible with $(4, 2)$ is $4(4) = 8$, and the number of ranked tree shapes compatible with $(4, 2, 6)$ is $\left(\frac{10}{5}\right)(|G(4, 2)| |G(6)| - \frac{1}{2} |G(4, 2)|^2) = \left(\frac{10}{5}\right)(8)(16) - \frac{1}{2}(8)^2 = 24,192$.

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 the numbers of ranked 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 compatible with the perfect phylogenies, just as the numbers of ranked unlabeled tree shapes exceed the numbers of unranked unlabeled tree shapes (Sect. 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 $b_1(n), b_2(n), b_3(n), \ldots$, 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 12 Among the integer sequences $b_1(n), b_2(n), b_3(n), \ldots$, 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 $\beta_x$ 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 $\beta_x$ cannot have the same sample size, we apply Proposition 11, finding
that the number of ranked unlabeled tree shapes compatible with $\beta_x$ is

$$\left(\frac{n}{x}\right) e_x e_{n-x}, \quad (10)$$

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

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

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

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

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

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

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

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 $\beta_x$ behaves asymptotically like the product of $n!$ and

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

where

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

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

$$c_x = 1 + \left(-\frac{1}{3^{x+1}} + \frac{1}{5^{x+1}}\right) + \left(-\frac{1}{7^{x+1}} + \frac{1}{9^{x+1}}\right) + \ldots \leq 1. \quad (16)$$
Among odd values of $x$, we have $c_1 = \frac{3}{2} \zeta(2) = \frac{\pi^2}{8} \approx 1.2337$ for $x = 1$. For odd $x \geq 3$, we have

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

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

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 $\beta_x$ 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 12. 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

$$b_2(n) \sim 2 \left(\frac{2}{\pi}\right)^{n-2} (n-2)!, \quad (17)$$

which, for $n = 10$, yields $b_2(10) \approx 20,643,840/\pi^8 \approx 2175.66$.

We also obtain the following corollary.

**Corollary 13** Among the integer sequences $b_1(n)$, $b_2(n)$, $b_3(n)$, . . . , the sequence $b_1(n)$ has the slowest asymptotic growth.

**Proof** From the proof of Proposition 12, for $x \geq 0$, $b_{x+1}(n+2)$ gives the number of ranked unlabeled shapes compatible with $\beta_x = (x + 1, n - x + 1)$. The proof obtains $b_{x+1}(n + 2) = n! q_x q_{n-x}$, or, following Eq. 14, $b_{x+1}(n + 2) \sim 4(\frac{2}{\pi})^{n+2} c_x n!$, where $c_x$ follows Eq. 15. Hence, to show $b_1(n)$ has the slowest growth among $b_1(n)$, $b_2(n)$, $b_3(n)$, . . . , it suffices to show that in Eq. 15, among all values of $x \geq 0$, $c_0$ is the smallest.

We see that $c_0$ is equal to the power series expansion of $\arctan(1)$, or $\pi/4$. For even $x \geq 2$, consider the expansion of $c_x$ in Eq. 16, and let $f(x, A) = -1/A^{x+1} + 1/(A + 2)^{x+1}$, so that $c_x = 1 + f(x, 3) + f(x, 7) + f(x, 11) + \ldots$. We claim that termwise, for even $x \geq 2$ and $A \geq 3$, $f(x, A) > f(0, A)$, so that summing terms in Eq. 16, we obtain $c_x > c_0$ for even $x \geq 2$.

To prove the claim, it suffices to show that for fixed $A \geq 3$, $f$ increases from $x = 0$, or $\frac{\partial f(x, A)}{\partial x} > 0$ for $x > 0$. We have $\frac{\partial f(x, A)}{\partial x} = \frac{\log A}{A^{x+1} - [\log(A + 2)]/(A + 2)^{x+1}}$. To verify that $\frac{\partial f(x, A)}{\partial x} > 0$ for $x > 0$ and $A \geq 3$, we see that $\frac{\partial f(x, A)}{\partial x} > 0$ is equivalent to $(\frac{A+2}{A})^{x+1} > \frac{\log(A+2)}{\log A}$. Now, from the inequality $1 + y \leq e^y$, we obtain $1 + \frac{2}{A} \leq e^{2/A}$ and hence $(\frac{A+2}{A})^A \leq e^2$. Because $A \geq 3$, $A^2 > e^2$. Therefore, $A^2 > (\frac{A+2}{A})^A$, from which $A^{A+2} > (A + 2)^A$, $(A + 2) \log A > A \log(A + 2)$, and $\frac{A+2}{A} > \frac{\log(A+2)}{\log A}$. We then have $(\frac{A+2}{A})^{x+1} > \frac{A+2}{A} > \frac{\log(A+2)}{\log A}$, and $\frac{\partial f(x, A)}{\partial x} > 0$.

We conclude $c_x > c_0$ for all even $x \geq 2$. From the proof of Proposition 12, we know $c_1 = \pi^2/8 > \pi/4 = c_0$. For odd $x \geq 3$, in Eq. 15, $\sum_{k=1}^{\infty} 1/k^{x+1} > 1$, so that $c_x > 1 - 1/2^{x+1} \geq 1 - 1/2^{3+1} = \frac{15}{16} > \pi/4 = c_0$, completing the proof.
The proof of Corollary 13 yields
\[ b_1(n) \sim 2^{\frac{2}{\pi}} (n - 2)! \text{.} \] (18)

The approximation yields \( b_1(10) \approx 41,287,680/\pi^9 \approx 1385.07 \), and we see in Table 1 that \( b_1(10) = 1385 \). From Eqs. 17 and 18, we have \( b_2(n)/b_1(n) \sim \pi/2 \), and we see from Table 1 that \( b_2(10)/b_1(10) = 2176/1385 \approx 1.57 \approx \pi/2 \).

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

Propositions 7, 8, 10 and 11 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 Sect. 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 Fig. 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 14** (Ranked labeled tree shape \( T_n^L \) compatible with a labeled perfect phylogeny \( \pi \in \Pi_L^L \)). A ranked labeled tree shape with \( n \) leaves, \( T_n^L \), is compatible with a perfect phylogeny \( \pi \in \Pi_L^L \) if the unranked unlabeled tree shape \( T_n \) obtained by removing the ranks and the labels from \( T_n^L \) is compatible with \( \pi \) and the one-to-one correspondence between the \( k \) leaves of \( \pi \) and the \( k \) disjoint subtrees of \( T_n^L \) correspond to the same partition of the individual sequences.

**Proposition 15** 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
\[ |G_c^L(\pi)| = \binom{n_1 + n_2 - 2}{n_1 - 1} |G_c^L(\pi_1)| |G_c^L(\pi_2)|. \] (19)

**Proof** We can count the number of ranked labeled tree shapes by dividing \( \pi \) at the root into two subtrees, one with \( n_1 \) leaves and perfect phylogeny \( \pi_1 \), and the other with \( n_2 \) leaves and perfect phylogeny \( \pi_2 \), 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 $\pi_1$ and $\pi_2$ can never be identical because they correspond to different sets of sequences.

Counts for the number of ranked labeled tree shapes for some of the perfect phylogenies of 10 taxa (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 the labeled perfect phylogenies from Sect. 3.4, we now count the unranked labeled tree shapes compatible with a labeled binary perfect phylogeny.

Proposition 16. Fix $n \geq 2$. Among the values $b_1'(n), b_2'(n), \ldots, b_{\lfloor n/2 \rfloor}'(n)$, the largest is $b_1'(n)$, and the smallest is $b_{\lfloor n/2 \rfloor}'(n)$.

Proof. Applying Proposition 15, 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 $x = \lfloor n/2 \rfloor$, $b_x'$ decreases monotonically from $x = 1$ to $x = \lfloor n/2 \rfloor$.

An example of Proposition 16 is visible in Table 1, in which $b_1'(10) = 57,153,600$ exceeds $b_2'(10), b_3'(10), b_4'(10)$, and $b_5'(10)$, among which $b_5'(10) = 2,268,000$ is the smallest.

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

Continuing with the labeled perfect phylogenies from Sect. 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 Fig. 8B. We calculate the number of unranked labeled tree shapes compatible with a labeled binary perfect phylogeny. Each row of Fig. 8C corresponds to one of the unranked labeled tree shapes compatible with the labeled perfect phylogeny.

Definition 17. (Unranked 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 $\pi$ and the one-to-one correspondence between the $k$ leaves of $\pi$ and the $k$ disjoint subtrees of $T_n^X$ correspond to the same partition of the individual sequences.
Fig. 8  Coalescent and infinitely-many-sites generative model of binary molecular data.  

**A**  Labeled genealogy with mutations  

| 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**  Labeled perfect phylogeny  

| Perfect phylogeny | Topology only |
|--------------------|---------------|
| b                  | a             |
| 2                  | 3             |
| 2                  | 3             |

Data: two haplotypes are observed, 01 and 10, grouping individuals $\{x_1, x_2\}$ and $\{x_3, x_4, x_5\}$, respectively.

**C**  Possible labeled ranked tree shapes compatible with perfect phylogeny  

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)
Proposition 18 For \( n_1, n_2 \geq 1 \) and \( \pi_1 \in \prod_{n_1}^L, \pi_2 \in \prod_{n_2}^L \), the number of unranked labeled tree shapes compatible with a labeled binary perfect phylogeny \( \pi = (\pi_1, \pi_2) \) is

\[
|G_c^X(\pi)| = |G_c^X(\pi_1)| |G_c^X(\pi_2)|.
\]  

**Proof** We divide \( \pi \) at the root into two subtrees, one with \( n_1 \) leaves and perfect phylogeny \( \pi_1 \), and the other with \( n_2 \) leaves and perfect phylogeny \( \pi_2 \). The subtrees must partition the sampled sequences in the same way as \( \pi \). The number of such trees is the simply product of the numbers of unranked labeled trees for the two subtrees. As in Proposition 15, perfect phylogenies \( \pi_1 \) and \( \pi_2 \) are not identical because they correspond to different sets of sequences; with the ranking dropped, unlike in Proposition 15, we need not consider the number of ways of interleaving the internal nodes of the two subtrees. \( \square \)

For some of the perfect phylogenies of 10 taxa (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 \( |G_c^X((4,2))| |G_c^X((6))| = |G_c^X((4))| |G_c^X((2))| |G_c^X((6))| = X_4X_2X_6 = 15 \times 1 \times 945 = 14,175 \).

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 16. Let \( b_c^x(n) \) denote the number of unranked labeled tree shapes compatible with \( B_x(n) \).

Proposition 19 Fix \( n \geq 2 \). Among the values \( b_c^x(n), b_c^x(n), \ldots, b_c^{\frac{n}{2}}(n) \), the largest is \( b_c^x(n) \), and the smallest is \( b_c^{\frac{n}{2}}(n) \).

**Proof** Applying Proposition 18, we have \( b_c^n(n) = X_n X_{n-x} \) for \( 1 \leq x \leq \lfloor \frac{n}{2} \rfloor \). Simplifying with Eq. 2, we obtain

\[
b_c^n(n) = \frac{(n-2)!}{2^{n-2}} \cdot \frac{\binom{2x-2}{x-1}}{\binom{n-2}{x-1}}.
\]

Then \( b_c^n(n)/b_c^{n+1}(n) = \frac{2x-1}{2n-2x-3} \leq 1 \) for \( 1 \leq x \leq \frac{n-1}{2} \), with equality requiring \( x = \frac{n-1}{2} \), so that \( b_c^n(n) \) monotonically decreases from \( x = 1 \) to \( x = \lfloor \frac{n}{2} \rfloor \).

In Table 1, we observe that as in Proposition 19, \( b_c^n(10) = 2,027,025 \) exceeds \( b_c^2(10) \), \( b_c^3(10) \), \( b_c^4(10) \), and \( b_c^n(10) \), among which \( b_c^4(10) = 11,025 \) is the smallest.

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 Fig. 9. Because multiple leaves can each correspond to groups with the same number of taxa, 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 Fig. 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 \( \sigma \) can be obtained by sequentially collapsing pairs of pendant edges of \( \pi \). Given two multifurcating perfect phylogenies \( \pi_1 \) and \( \pi_2 \), their meet \( \pi_1 \land \pi_2 \) is the largest multifurcating perfect phylogeny that refines both \( \pi_1 \) and \( \pi_2 \). For example, the meet between \( \pi_1 = (1, 2, 3, (2, 2)) \) and \( \pi_2 = (1, 2, 2, (2, 3)) \) is given by:

\[
(1, 2, 3, (2, 2)) \land (1, 2, 2, (2, 3)) = (1, 2, 2, (2, 3)).
\]

Similarly, their join is the smallest multifurcating perfect phylogeny \( \pi_1 \lor \pi_2 \) for which both \( \pi_1 \) and \( \pi_2 \) are refinements:

\[
(1, 2, 3, (2, 2)) \lor (1, 2, 2, (2, 3)) = (1, 2, 2, 2, 3).
\]

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 10 and 11 to each binary perfect phylogeny. To recursively generate all possible binary perfect phylogenies from \( \pi \), we define the operator \( B_{i,j}(\pi) \) that collapses two leaves with labels \( a_i \) and \( a_j \) in \( \pi \). For example, \( B_{2,3}(2, 2, 3, 4) = (2, 3, 2, 4) \). If \( \sum_{i=1}^{s} m_i > 2 \), then

\[
|G_c(\pi)| = \sum_{i=1}^{s} |G_c(B_{i,i}(\pi))| 1_{m_i > 1} + \sum_{i=1}^{s-1} \sum_{j=i+1}^{s} |G_c(B_{i,j}(\pi))| 1_{m_i > 0} 1_{m_j > 0}
\]

- collapsing two pendant edges with the same leaf values
- collapsing two pendant edges with different leaf values
- collapsing all pairs containing two distinct pairs of pendant edges, each pair with the same leaf values
The equation proceeds until all terms reach \( \sum \) the calculation in terms of binary perfect phylogenies. The recursive application of beled tree shapes compatible with a multifurcating perfect phylogeny by expressing Example 20

The number of ranked unlabeled tree shapes compatible with phylogenies are reached. in place of Propositions 10 and 11.

Equation 21 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 \( \sum_{i=1}^{\infty} m_i = 2 \), when the binary perfect phylogenies are reached.

Example 20 The number of ranked unlabeled tree shapes compatible with \( \pi = (2, 2, 3, 3) \) is:

\[
|G_c^T (2, 2, 3, 3)| = |G_c^T ((2, 2), (3, 3))| + |G_c^T (2, 2, (3, 3))| \\
+ |G_c^T ((2, 3), (2, 3))| - |G_c^T ((2, 2), (3, 3))| \\
= \left[ |G_c^T ((2, 2), (3, 3))| + |G_c^T ((2, 2), (3, 3))| \right] \\
+ \left[ |G_c^T ((2, 2), (3, 3))| + |G_c^T ((3, 3), (2, 2))| \right] \\
+ \left[ |G_c^T ((2, 3), (2, 3))| + |G_c^T ((2, 3), (2, 3))| \right] \\
- |G_c^T ((2, 2), (3, 3))| \\
= |G_c^T ((2, 2), (3, 3))| + |G_c^T ((2, 2), (3, 3))| \\
+ |G_c^T ((3, 3), (2, 2))| + |G_c^T ((2, 3), (2, 3))| \\
+ |G_c^T ((2, 2), (3, 3))| + |G_c^T ((2, 3), (2, 3))| \\
= 168 + 280 + 144 + 420 + 360 + 315 = 1687.
\]

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

For counting the number of unranked unlabeled tree shapes compatible with \( \pi = (n_1, n_2, \ldots, n_s) \), we simply replace \( G_c^T \) with \( G_c \) in Eq. 21. We use Propositions 7 and 8 in place of Propositions 10 and 11.
Example 21 The number of unranked unlabeled tree shapes compatible with $\pi = (2, 2, 3, 3)$ is:

$$|G_c(2, 2, 3, 3)| = |G_c((2, 2), (3, 3))| + |G_c(((2, 2), 3), (3, 3))| + |G_c(((3, 3), 2), (2, 2))| + |G_c(((2, 2), 3), 2)| + |G_c((2, 3), 2)| + |G_c((2, 3), (2, 3))| = 1 + 1 + 1 + 1 + 1 + 1 = 6.$$ 

This example is quite straightforward; the values of $G_c$ for the perfect phylogenies that appear in intermediate steps—$(2)$, $(3)$, $(2, 2)$, $(3, 3)$, $(2, 3)$, $(2, 2, 3)$, $(2, 3, 2)$, and $(2, 3, 2)$—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 taxa, all of which are all assumed to have different labels. Therefore, we take $a = (n_1, n_2, \ldots, n_k)$ and $m = (1, 1, \ldots, 1)$. Equation 21 reduces to

$$|G_c^L(\pi)| = \sum_{i=1}^{s-1} \sum_{j=i+1}^{s} |G_c^L(B_{i, j}(\pi))| l_{m_i} > 0 l_{m_j} > 0$$

(collapsing two pendant edges)

$$- \sum_{i=1}^{s-1} \sum_{j=i+1}^{s} \sum_{k=1}^{s} \sum_{\ell=k+1}^{s} |G_c^L(B_{i, j}(\pi) \wedge B_{k, \ell}(\pi))| l_{m_i} > 0 l_{m_j} > 0 l_{m_k} > 0 l_{m_\ell} > 0. \quad (22)$$

(collapsing two pairs of pendant edges)

The enumeration makes use of Proposition 15.

Example 22 Consider a labeled multifurcating perfect phylogeny that groups 2, 2, 3, and 3 taxa at the root. We assume that $a = (a_1, a_2, a_3, a_4) = (2, 2, 3, 3)$. Applying the recursion formula in Eq. 22, we get

$$|G_c^L(a_1, a_2, a_3, a_4)|$$

$$= |G_c^L((a_1, a_2), a_3, a_4)| + |G_c^L((a_1, a_3), a_2, a_4)| + |G_c^L((a_1, a_4), a_2, a_3)|$$

$$+ |G_c^L((a_2, a_3), a_1, a_4)| + |G_c^L((a_2, a_4), a_1, a_3)| + |G_c^L((a_3, a_4), a_1, a_2)|$$

$$- |G_c^L((a_1, a_2), (a_3, a_4))| - |G_c^L((a_1, a_3), (a_2, a_4))| - |G_c^L((a_1, a_4), (a_2, a_3))|$$

$$= |G_c^L((2, 2), 3, 3)| + 4|G_c^L((2, 3), 2, 3)| + |G_c^L((3, 3), 2, 2)|$$

$$- |G_c^L((2, 2), (3, 3))| - 2|G_c^L((2, 3), (2, 3))|.$$ 

Now, because

$$|G_c^L(a_1, a_2, a_3)| = |G_c^L((a_1, a_2), a_3)| + |G_c^L((a_1, a_3), a_2)| + |G_c^L((a_2, a_3), a_1)|.$$
we have

\[ |G_c^L((2, 2), 3, 3)| = 2|G_c^L(((2, 2), 3), 3)| + |G_c^L((2, 2), (3, 3))| \]
\[ |G_c^L((2, 3), 2, 3)| = |G_c^L(((2, 3), 2), 3)| + |G_c^L(((2, 3), 3), 2)| + |G_c^L((2, 3), (2, 3))| \]
\[ |G_c^L((3, 3), 2, 2)| = 2|G_c^L(((3, 3), 2), 2)| + |G_c^L((2, 2), (3, 3))|. \]

Summing all terms, we get

\[ |G_c^L(a_1, a_2, a_3, a_4)| \]
\[ = 2|G_c^L(((2, 2), 3), 3)| + 2|G_c^L(((3, 3), 2), 2)| + |G_c^L((2, 2), (3, 3))| \]
\[ + 4|G_c^L(((2, 3), 2), 3)| + 4|G_c^L(((2, 3), 3), 2)| + 2|G_c^L((2, 3), (2, 3))| \]
\[ = 2 \times 5040 + 2 \times 2592 + 6048 + 4 \times 3780 + 4 \times 3240 + 2 \times 5670 = 60,732. \]

In obtaining this sum, we use the fact that the values of \( G_c^L \) 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 \( G_c^L \) with \( G_c^X \) in Eq. 22. We use Proposition 18 in place of Proposition 15.

**Example 23** The number of unranked labeled tree shapes compatible with a labeled multifurcating perfect phylogeny that groups 2, 2, 3, and 3 taxa at the root, with \( a = (a_1, a_2, a_3, a_4) = (2, 2, 3, 3) \) is:

\[ |G_c^X(a_1, a_2, a_3, a_4)| \]
\[ = 2|G_c^X(((2, 2), 3), 3)| + 2|G_c^X(((3, 3), 2), 2)| + |G_c^X((2, 2), (3, 3))| \]
\[ + 4|G_c^X(((2, 3), 2), 3)| + 4|G_c^X(((2, 3), 3), 2)| + 2|G_c^X((2, 3), (2, 3))| \]
\[ = 2 \times 9 + 2 \times 9 + 9 + 4 \times 9 + 4 \times 9 + 2 \times 9 = 135. \]

The sum uses values of \( G_c^X \) 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.

The entries in the table are obtained by repeated use of Propositions 7 and 8 for unranked unlabeled tree shapes, Propositions 10 and 11 for ranked unlabeled tree shapes, Proposition 15 for ranked labeled tree shapes, and Proposition 18 for unranked labeled tree shapes. An arbitrary labeling of the perfect phylogeny is assumed for counting the associated ranked and unranked labeled tree shapes. Figure 10 shows the corresponding partial Hasse diagram of the lattice of binary perfect phylogenies with 10 taxa.
Fig. 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.

Table 1 Number of trees compatible with example perfect phylogenies of 10 taxa

| 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                        |
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 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 5). 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 perfect phylogenies. Further, for multifurcating perfect phylogenies, we have exploited a recursive inclusion-exclusion principle to decompose a multifurcating perfect phylogeny into all possible binary perfect phylogenies, extending the utility of our lattice approach from bifurcating structures to more general structures.

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 Sects. 3.2, 3.3, 3.4, 3.5, and 4. We have provided examples for relatively small cases with $n = 10$ taxa (Table 1, Fig. 10). 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. Our algorithmic implementation in Python does have a computational precision limitation, but it accommodates numbers up to the order of $10^{290}$.

We obtained results concerning the cherry perfect phylogenies with the largest numbers of ranked unlabeled, unranked labeled, and ranked labeled tree shapes (Propositions 12, 16, and 19), and it will be informative to seek a similar result for the unranked unlabeled case. The result in Proposition 12 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 12, 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 1993; 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 (Alimpiev and Rosenberg 2022); 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 (and unranked) unlabeled tree shapes compatible with an observed perfect phylogeny, relative to the number of ranked (and unranked) labeled tree shapes compatible with an observed perfect phylogeny (Tables 1 and 2). This observation contributes to a growing branch of the area of coalescent-based inference (Sainudin...
Table 2  Ratio of the number of unranked labeled and unranked unlabeled tree shapes and ratio of the number of ranked labeled and ranked unlabeled tree shapes compatible with three perfect phylogenies of 10, 20 and 50 taxa

| Perfect phylogeny | Ratio for unranked tree shapes | Ratio for ranked tree shapes |
|-------------------|---------------------------------|-----------------------------|
| (5,5)             | 1837.5                          | 2592                        |
| (15,5)            | $\sim 1.54 \times 10^{12}$      | $\sim 1.26 \times 10^{12}$  |
| (45,5)            | $\sim 2.17 \times 10^{53}$      | $\sim 8.18 \times 10^{52}$  |

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.

Acknowledgements J.A.P. and N.A.R. acknowledge support from National Institutes of Health Grant R01-GM-131404. J.A.P. acknowledges support from the Alfred P. Sloan Foundation.

Appendix: Proof of Theorem 5

To prove Theorem 5, 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 $(\pi_1, \pi_2)$ 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 $(\Pi_k \cup \{\emptyset\}, \land, \lor)$ is a lattice for all $k, 1 \leq k < n$. We then verify that it follows that $(\Pi_n \cup \{\emptyset\}, \land, \lor)$ is a lattice. We start with Condition 2, which is trivial.

Condition 2: $\pi \land \sigma = \sigma \land \pi$ and $\pi \lor \sigma = \sigma \lor \pi$

For all $n$, condition 2 of the definition of a lattice is trivially satisfied, as the operations $\land$ and $\lor$ 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 $\Pi_1$ 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 $(1) \land (1) = (1)$, $\emptyset \land \emptyset = \emptyset$, $(1) \lor (1) = (1)$, and
\(\emptyset \lor \emptyset = \emptyset\). These four relations are true by parts (3), (1), (4), and (2) of Definition 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 Definition 1 (1–4). Hence, \((\Pi_1 \cup \{\emptyset\}, \land, \lor)\) is a lattice.

**Condition 1:** \(\pi \land \pi = \pi\) and \(\pi \lor \pi = \pi\)

First, we demonstrate the first part of the condition. We see \(\emptyset \land \emptyset = \emptyset\) by Definition 1 (1) and \((\pi) \land (\pi) = (\pi)\) by Definition 1 (3).

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

\[
\pi \land \pi = (\pi_1, \pi_2) \land (\pi_1, \pi_2) \\
= (\pi_1 \land \pi_1, \pi_2 \land \pi_2) \lor (\pi_1 \land \pi_2, \pi_2 \land \pi_1) \quad \text{by Prop. 2 (2)} \\
= (\pi_1, \pi_2) \lor (\pi_1 \land \pi_2, \pi_1 \land \pi_2) \quad \text{by the inductive hypothesis.}
\]

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

If \(\pi_1 \leq \pi_2\), then \(\pi_1 \land \pi_2 = \pi_1\) and \(\pi_1 \lor \pi_2 = \pi_2\). By Proposition 2 (3), \((\pi_1, \pi_2) \lor (\pi_1, \pi_1) = (\pi_1, \pi_2 \lor \pi_1) = (\pi_1, \pi_2) = \pi\), so that \(\pi \land \pi = \pi\).

If \(\pi_1\) and \(\pi_2\) are not comparable, then by Definition 1 (11), \(\pi_1 \land \pi_2 = \delta\) for some \(\delta \in (\Pi_{n_1} \cup \{\emptyset\}) \setminus \{\pi_1, \pi_2\}\), with \(\delta \lor \pi_1 = \pi_1\) and \(\delta \lor \pi_2 = \pi_2\). We then have by Proposition 2 (3),

\[
(\pi_1, \pi_2) \lor (\pi_1 \land \pi_2, \pi_1 \land \pi_2) = (\pi_1, \pi_2) \lor (\delta, \delta).
\]

But \((\delta, \delta)\) refines \((\pi_1, \pi_2)\), as \(\delta\) refines \(\pi_1\) and \(\delta\) refines \(\pi_2\), so that \((\pi_1, \pi_2)\) can be obtained by collapsing cherries separately in the two subtrees of \((\delta, \delta)\). Hence, \(\pi \land \pi = (\pi_1, \pi_2) \lor (\delta, \delta) = (\pi_1, \pi_2) = \pi\).

For the second part of the condition, we have \(\emptyset \lor \emptyset = \emptyset\) by Definition 1 (2) and \((\pi) \lor (\pi) = (\pi)\) by Definition 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 \(n_1 + n_2 = n\).

\[
\pi \lor \pi = (\pi_1, \pi_2) \lor (\pi_1, \pi_2) \\
= (\pi_1, \pi_2 \lor \pi_2) \quad \text{by Prop. 2 (3)} \\
= (\pi_1, \pi_2) \quad \text{by the inductive hypothesis} \\
= \pi.
\]
**Condition 4:** \(\pi \land (\pi \lor \sigma) = \pi\) and \(\pi \lor (\pi \land \sigma) = \pi\)

First, we see that both parts of the condition hold if at least one of \(\pi, \sigma\) is in \(\{\emptyset, (n)\}\), by Definition 1 (1–4). Next, we have the following 3 cases:

i. If \(\pi \leq \sigma\), then \(\pi \land \sigma = \pi\) and \(\pi \lor \sigma = \sigma\). Hence, \(\pi \land (\pi \lor \sigma) = \pi \land \sigma = \pi\).

By Condition 1, \(\pi \lor (\pi \land \sigma) = \pi \lor \pi = \pi\).

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

iii. If \(\pi\) and \(\sigma\) are not comparable, then by Proposition 2 (5), there exists a perfect phylogeny \(\gamma\) such that \(\pi \lor \sigma = \gamma\), \(\pi \land \gamma = \pi\), and \(\sigma \land \gamma = \sigma\). Hence \(\pi \land (\pi \lor \sigma) = \pi \land \gamma = \pi\). By Proposition 2 (5), there exists a perfect phylogeny \(\rho\) such that \(\pi \land \sigma = \rho\), \(\pi \lor \sigma = \pi\), and \(\sigma \lor \rho = \sigma\). We have \(\pi \lor (\pi \land \sigma) = \pi \lor \rho = \pi\).

**Condition 3:** \(\pi \land (\sigma \land \rho) = (\pi \land \sigma) \land \rho\) and \(\pi \lor (\sigma \lor \rho) = (\pi \lor \sigma) \lor \rho\)

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

\[
\pi \land (\sigma \land \rho) \\
= (\pi_1, \pi_2) \land ((\sigma_1, \sigma_2) \land (\rho_1, \rho_2)) \\
= (\pi_1, \pi_2) \land [(\sigma_1 \land \rho_1, \sigma_2 \land \rho_2) \lor (\sigma_1 \land \rho_2, \sigma_2 \land \rho_1)] \text{ by Prop. 2 (2)} \\
= [(\pi_1, \pi_2) \land (\sigma_1 \land \rho_1, \sigma_2 \land \rho_2)] \lor [(\pi_1, \pi_2) \land (\sigma_1 \land \rho_2, \sigma_2 \land \rho_1)] \text{ by Prop. 2 (4)} \\
= [(\pi_1 \land (\sigma_1 \land \rho_1), \pi_2 \land (\sigma_2 \land \rho_2)) \lor (\pi_1 \land (\sigma_2 \land \rho_2), \pi_2 \land (\sigma_1 \land \rho_1))] \\
\lor [(\pi_1 \land (\sigma_1 \land \rho_2), \pi_2 \land (\sigma_2 \land \rho_1)) \lor (\pi_1 \land (\sigma_2 \land \rho_1), \pi_2 \land (\sigma_1 \land \rho_2))] \text{ by Prop. 2 (2)}
\]

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

\[
\pi \land (\sigma \land \rho) \equiv [(\pi_1 \land \sigma_1) \land \rho_1, (\pi_2 \land \sigma_2) \land \rho_2) \lor (\pi_1 \land \sigma_2) \land \rho_2, (\pi_2 \land \sigma_1) \land \rho_1)] \\
\lor [(\pi_1 \land \sigma_1) \land \rho_2, (\pi_2 \land \sigma_2) \land \rho_1) \lor (\pi_1 \land \sigma_2) \land \rho_1, (\pi_2 \land \sigma_1) \land \rho_2)].
\]

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

\[
\pi \land (\sigma \land \rho) = ((\pi_1 \land \sigma_1) \land \rho_1, (\pi_2 \land \sigma_2) \land \rho_2) \lor [(\pi_1 \land \sigma_1) \land \rho_2, (\pi_2 \land \sigma_2) \land \rho_1) \\
\lor ((\pi_1 \land \sigma_2) \land \rho_2, (\pi_2 \land \sigma_1) \land \rho_1)] \lor ((\pi_1 \land \sigma_2) \land \rho_1, (\pi_2 \land \sigma_2) \land \rho_2)].
\]

Dropping the brackets and viewing this expression as having four perfect phylogenies separated by the \(\lor\) operator, we group the first two and the last two perfect phylogenies together and apply Proposition 2 (2) to each group. We get
\[ \pi \land (\sigma \land \rho) = [(\pi_1 \land \sigma_1, \pi_2 \land \sigma_2) \land (\rho_1, \rho_2)] \]
\[ \lor [(\pi_1 \land \sigma_2, \pi_2 \land \sigma_1) \land (\rho_1, \rho_2)] \text{ by Prop. 2 (2)} \]
\[ = [(\pi_1 \land \sigma_1, \pi_2 \land \sigma_2) \lor (\pi_1 \land \sigma_2, \pi_2 \land \sigma_1)] \land (\rho_1, \rho_2) \text{ by Prop. 2 (4)} \]
\[ = (\pi \land \sigma) \land \rho \text{ by Prop. 2 (2)}. \]

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 \( \pi \lor \sigma = \pi \lor (\sigma \land \rho) = \pi \lor (\rho \land \rho) = \pi \lor (n) = \pi \lor (n) = (n) = (\pi) \land (\sigma) \land \rho \text{ by Definition 1 (4)}, \)
\[ \text{Note that this derivation includes the case of shared subtrees at the root, in which it is not only the sizes of the subtrees that are the same, but the subtrees themselves. For example, suppose } \pi = (\pi_1, \pi_2) \text{ and } \sigma = (\pi_1, \sigma_1). \text{ By Proposition 2 (3), we have} \]
\[ \pi \lor \sigma = (\pi_1, \pi_2) \lor (\pi_1, \sigma_1) = (\pi_1, \pi_2 \lor \sigma_1). \]
However, we will show that we can replace the previous equality by the extended expression:

$$\pi \lor \sigma = (\pi_1 \lor \pi_1, \pi_2 \lor \sigma_1) \land (\pi_1 \lor \sigma_1, \pi_1 \lor \pi_2), \quad (24)$$

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

$$\pi_2 \lor \sigma_1 = (\pi_2 \lor (\pi_1 \land \pi_2)) \lor \sigma_1 \quad \text{by Condition 4}$$
$$= (\pi_2 \lor \sigma_1) \lor (\pi_1 \land \pi_2) \quad \text{by the inductive hypothesis and Condition 2}$$
$$= \pi_2 \lor [(\sigma_1 \land \pi_1) \lor \sigma_1] \lor (\pi_1 \land \pi_2) \quad \text{by Conditions 2 and 4}$$
$$= [\pi_2 \lor (\sigma_1 \land \pi_1)] \lor [\sigma_1 \lor (\pi_1 \land \pi_2)] \quad \text{by the inductive hypothesis.}$$

Then

$$\pi \lor \sigma = (\pi_1, \pi_2 \lor \sigma_1)$$
$$= (\pi_1, [\pi_2 \lor (\sigma_1 \land \pi_1)] \lor [\sigma_1 \lor (\pi_1 \land \pi_2)])$$
$$= (\pi_1, \pi_2 \lor (\sigma_1 \land \pi_1)) \lor (\pi_1, \sigma_1 \lor (\pi_1 \land \pi_2)) \quad \text{by Prop. 2 (3)}$$
$$= (\pi_1, (\pi_2 \lor \sigma_1) \lor (\pi_2 \lor \pi_1)) \lor (\pi_1, (\sigma_1 \lor \pi_1) \land (\sigma_1 \lor \pi_2)) \quad \text{by Prop. 2 (4).} \quad (25)$$

By Condition 4 and Proposition 2 (3) we have

$$\pi_1 = \pi_1 \lor (\pi_1 \land \sigma_1) = (\pi_1 \lor \pi_1) \land (\pi_1 \lor \sigma_1),$$

and

$$\pi_1 = \pi_1 \lor (\pi_1 \land \pi_2) = (\pi_1 \lor \pi_1) \land (\pi_1 \lor \pi_2).$$

Replacing the first $\pi_1$ in the first pair of Eq. 25 by $(\pi_1 \lor \pi_1) \land (\pi_1 \lor \sigma_1)$, and the first $\pi_1$ in the second pair of Eq. 25 by $(\pi_1 \lor \pi_1) \land (\pi_1 \lor \pi_2)$, we get

$$\pi \lor \sigma = ((\pi_1 \lor \pi_1) \land (\pi_1 \lor \sigma_1), (\pi_2 \lor \sigma_1) \land (\pi_2 \lor \pi_1)) \lor ((\pi_1 \lor \pi_1)$$
$$\land (\pi_1 \lor \pi_2), (\sigma_1 \lor \pi_1) \land (\sigma_1 \lor \pi_2))$$
$$= (\pi_1 \lor \pi_1, \pi_2 \lor \sigma_1) \land (\pi_1 \lor \sigma_1, \pi_1 \lor \pi_2) \quad \text{by Prop. 2 (2).}$$

Thus, Eq. 24 holds, so that Eq. 23 holds for the case in which subtrees are shared at the root.

**Appendix: Proof of Proposition 2**

1. Let $\pi_1 = (n_1, n_2)$ and $\pi_2 = (n_3, n_4)$ be two perfect phylogenies in $\Pi_n$ with $n_1 + n_2 = n_3 + n_4 = n$. Then
\[ \pi_1 \lor \pi_2 = (n_1, n_2) \lor (n_3, n_4) = \begin{cases} (n_1, n_2) & \text{if } n_1 = n_3 \text{ or } n_1 = n_4 \\ (n) & \text{otherwise.} \end{cases} \]

**Proof** (a) If \( n_1 = n_3 \) (or \( n_1 = n_4 \)), then \( n_2 = n_4 \) (or \( n_2 = n_3 \)) since \( n_1 + n_2 = n_3 + n_4 \) hence \( \pi_1 = \pi_2 \). This in turn implies that \( \pi_1 \leq \pi_2 \) and therefore \( \pi_1 \lor \pi_2 = \pi_2 = \pi_1 \).

(b) If \( n_1 \neq n_3 \) and \( n_1 \neq n_4 \), then by definition \( \pi_1 \lor \pi_2 \) is the smallest perfect phylogeny that is refined by both \( \pi_1 \) and \( \pi_2 \). Clearly \( (n) \) is refined by both \( \pi_1 \) and \( \pi_2 \) and this is the smallest since \( (n) \) is directly obtained by collapsing the pair of pendant edges of \( \pi_1 \) and \( \pi_2 \).

2. For all \( \pi_1, \pi_2, \pi_3, \pi_4 \) with \( (\pi_1, \pi_2) \in \Pi_n \) and \( (\pi_3, \pi_4) \in \Pi_n \),

\[ (\pi_1, \pi_2) \land (\pi_3, \pi_4) = (\pi_1 \land \pi_3, \pi_2 \land \pi_4) \lor (\pi_1 \land \pi_4, \pi_2 \land \pi_3), \]

with the convention that \( (\pi, \emptyset) = \emptyset \).

**Proof** The meet of the two perfect phylogenies is the largest perfect phylogeny that refines both \( (\pi_1, \pi_2) \) and \( (\pi_3, \pi_4) \), that is, the largest among \( (\pi_1 \land \pi_3, \pi_2 \land \pi_4) \) and \( (\pi_1 \land \pi_4, \pi_2 \land \pi_3) \), and the largest corresponds to their join.

3. For all \( \pi_1, \pi_2, \pi_3, \pi_4 \) with \( (\pi_1, \pi_2) \in \Pi_n \) and \( (\pi_3, \pi_4) \in \Pi_n \), \( \pi_i \in \Pi_n \), for \( i = 1, 2, 3, 4 \),

\[ (\pi_1, \pi_2) \lor (\pi_3, \pi_4) = \begin{cases} (n) & \text{if } n_1 \neq n_3 \text{ and } n_1 \neq n_4 \\ (\pi_1, \pi_2 \lor \pi_4) & \text{if } \pi_1 = \pi_3 \\ (\pi_1, \pi_2 \lor \pi_3) & \text{if } \pi_1 = \pi_4 \\ (\pi_2, \pi_1 \lor \pi_4) & \text{if } \pi_2 = \pi_3 \\ (\pi_2, \pi_1 \lor \pi_3) & \text{if } \pi_2 = \pi_4 \\ (\pi_1 \lor \pi_3, \pi_2 \lor \pi_4) \land (\pi_1 \lor \pi_4, \pi_2 \lor \pi_3) & \text{otherwise}, \end{cases} \]

with the convention that \( (\pi, \emptyset) = \emptyset \).

**Proof** The join of two perfect phylogenies is the smallest perfect phylogeny that is refined by both \( (\pi_1, \pi_2) \) and \( (\pi_3, \pi_4) \). (a) If \( n_1 \neq n_3 \) and \( n_1 \neq n_4 \), then no subtree of \( \pi_i \), \( i = 1, \ldots, 4 \) can be refined from any other by Definition 1 (6), therefore the join corresponds to the perfect phylogeny obtained by collapsing the pair of pendant edges of \( (\pi_1, \pi_2) \) and \( (\pi_3, \pi_4) \) which in both cases correspond to \( (n) \). (b) If one cross-pair of subtrees is identical, for example, \( \pi_1 = \pi_3 \), then clearly \( \pi_1 \lor \pi_3 = \pi_1 \) since \( \pi_1 \)
is refined by both $\pi_1$ and $\pi_3$, therefore the perfect phylogeny that is refined by both $(\pi_1, \pi_2)$ and $(\pi_1, \pi_4)$ is $(\pi_1, \pi_2 \lor \pi_4)$. (c) Otherwise, the join corresponds to the meet of the two perfect phylogenies that join two different cross-pairs from $(\pi_1, \pi_2)$ and $(\pi_3, \pi_4)$. □

4. For all $\pi_1, \pi_2, \pi_3 \in \Pi_n$,

$$\pi_1 \land (\pi_2 \lor \pi_3) = (\pi_1 \land \pi_2) \lor (\pi_1 \land \pi_3),$$

and

$$\pi_1 \lor (\pi_2 \land \pi_3) = (\pi_1 \lor \pi_2) \land (\pi_1 \lor \pi_3).$$

Proof In the first case, $\pi_1 \land (\pi_2 \lor \pi_3)$ is the largest perfect phylogeny that refines both $\pi_1$ and $(\pi_2, \pi_3)$. This perfect phylogeny then refines both $\pi_1$ and $\pi_2$ or both $\pi_1$ and $\pi_3$, that is, the largest of $(\pi_1, \pi_2)$ and $(\pi_1, \pi_3)$ and this largest perfect phylogeny corresponds to their join. In the second case, $\pi_1 \lor (\pi_2 \land \pi_3)$ is the smallest perfect phylogeny that is refined by both $\pi_1$ and $(\pi_2 \lor \pi_3)$, that is, refined by both $\pi_1$ and $\pi_2$, and by $\pi_1$ and $\pi_3$ and this corresponds to the meet of the perfect phylogeny that is refined by $\pi_1$ and $\pi_2$ and the perfect phylogeny that is refined by $\pi_1$ and $\pi_3$. □

5. 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

$$\pi \land \sigma = \gamma, \quad \pi \lor \gamma = \pi, \quad \text{and} \quad \sigma \lor \gamma = \sigma,$$

and

$$\pi \lor \sigma = \rho, \quad \pi \land \rho = \pi, \quad \text{and} \quad \sigma \land \rho = \sigma.$$

Proof The meet of two incomparable perfect phylogenies is the largest perfect phylogeny that refines both $\pi$ and $\sigma$. If $\pi$ or $\sigma$ or both are perfect phylogenies with all tips labeled 1, for example $((1,1),1)$ then the only refinement of $\pi$ and $\sigma$ is $\emptyset$ and the result holds by Definition 1. Otherwise a refinement of each perfect phylogeny can be obtained sequentially by branching any tip with label greater than 1, until a common perfect phylogeny is reached or until two perfect phylogenies with all tips labeled 1 are reached. If a common perfect phylogeny $\gamma$ is reached we then have $\gamma \lor \pi = \pi$ and $\gamma \lor \sigma = \sigma$ and the result holds. If no common perfect phylogeny is reached then $\gamma = \emptyset$ and the result holds. Similarly, the join of two incomparable perfect phylogenies is the smallest perfect phylogeny refined by both $\pi$ and $\sigma$. Since $\pi$ and $\sigma$ are not comparable then $\pi$ nor $\sigma$ are $(n)$, therefore we can sequentially collapse pairs of pendant edges of $\pi$ and $\sigma$ until a common perfect phylogeny is reached or $(n)$ is reached. In both cases, the result holds. □
References

Agarwala R, Fernández-Baca D (1993) A polynomial-time algorithm for the perfect phylogeny problem when the number of character states is fixed. SIAM J Comput 23(6):1216–1224

Alimpiev E, Rosenberg N (2022) A lattice structure for ancestral configurations arising from the relationship between gene trees and species trees. arXiv, arXiv:2111.10456v1 [q-bio.PE]

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

Brent RP, Harvey D (2013) Fast computation of Bernoulli, Tangent and Secant numbers. In: Computational and analytical mathematics. Springer, New York, pp 127–142

Cappello L, Palacios JA et al (2020a) Sequential importance sampling for multiresolution Kingman-Tajima coalescent counting. Ann Appl Stat 14(2):727–751

Cappello L, Yeber A, Palacios JA (2020b) The Tajima heterochronous n-coalescent: inference from heterochronously sampled molecular data. arXiv preprint arXiv:2004.06826

Disanto F, Rosenberg NA (2015) Coalescent histories for lodgepole species trees. J Comput Biol 22(10):918–929

Disanto F, Rosenberg NA (2017) Enumeration of ancestral configurations for matching gene trees and species trees. J Comput Biol 24(9):831–850

Drummond A, Suchard M, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969–1973

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland

Flajolet P, Sedgewick R (2009) Analytic combinatorics. Cambridge University Press, Cambridge

Griffiths RC, Tavaré S (1994) Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond Ser B Biol Sci 344(1310):403–410

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

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

Gusfield D (2014) ReCombinatorics: the algorithmics of ancestral recombination graphs and explicit phylogenetic networks. Cambridge

Harding EF (1971) The probabilities of rooted tree-shapes generated by random bifurcation. Adv Appl Probab 3(3):44–77

Janson S, Kersting G (2011) On the total external length of the Kingman coalescent. Electron J Probab 16:2203–2218

Jones MG, Khodaverdian A, Quinn JJ, Chan MM, Hussmann JA, Wang R, Xu C, Weissman JS, Yosef N (2020) Inference of single-cell phylogenies from lineage tracing data using Cassiopeia. Genome Biol 21:92

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

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

Kingman J (1982) The coalescent. Stoch Process Appl 13(3):235–248

Marjoram P, Tavaré S (2006) Modern computational approaches for analysing molecular genetic variation data. Nat Rev Genet 7(10):759–770

Nation JB (1998) Notes on lattice theory. http://www.math.hawaii.edu/~jb/lat1-6.pdf

Otter R (1948) The number of trees. Ann Math 49:583–599

Palacios JA, Wakeley J, Ramachandran S (2015) Bayesian nonparametric inference of population size changes from sequential genealogies. Genetics 201(1):281–304

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

Sainudiin R, Studler T, Véber A (2015) Finding the best resolution for the Kingman-Tajima coalescent: theory and applications. J Math Biol 70:1207–1247

Stanley RP (2012) Enumerative combinatorics, vol 1, 2nd edn. Cambridge studies in advanced mathematics. Cambridge University Press, New York

Steel M (2016) Phylogeny: discrete and random processes in evolution. SIAM, Philadelphia

Stephens M, Donnelly P (2000) Inference in molecular population genetics. J R Stat Soc Ser B (Stat Methodol) 62(4):605–635
Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105(2):437–460
Tavaré S (2004) Part I: ancestral inference in population genetics. Lectures on probability theory and statistics, volume 1837 of lecture notes in mathematics. Springer, New York, pp 1–188
Wakeley J (2008) Coalescent theory: an introduction. Roberts & Company Publishers, Greenwood Village
Wu Y (2012) Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evol Int J Org Evol 66(3):763–775

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.