Evolutionary Phylogenetic Networks: Models and Issues

Luay Nakhleh

Abstract Phylogenetic networks are special graphs that generalize phylogenetic trees to allow for modeling of non-treelike evolutionary histories. The ability to sequence multiple genetic markers from a set of organisms and the conflicting evolutionary signals that these markers provide in many cases, have propelled research and interest in phylogenetic networks to the forefront in computational phylogenetics. Nonetheless, the term ‘phylogenetic network’ has been generically used to refer to a class of models whose core shared property is tree generalization. Several excellent surveys of the different flavors of phylogenetic networks and methods for their reconstruction have been written recently. However, unlike these surveys, this chapter focuses specifically on one type of phylogenetic networks, namely *evolutionary* phylogenetic networks, which explicitly model reticulate evolutionary events. Further, this chapter focuses less on surveying existing tools, and addresses in more detail issues that are central to the accurate reconstruction of phylogenetic networks.

1 Introduction

In Charles Darwin’s *Origin of Species* [17], the depiction of an evolutionary history of species took the shape of a tree. Ever since, trees, in a variety of forms, have been the mainstream of phylogenetics. Such a tree, also referred to as a *phylogeny*, is taken to model the ancestor-descendant evolutionary relationship of a group of species from their most recent common ancestor. Though appropriate for several groups of taxa, a phylogenetic tree may be inadequate for other groups. For example, evidence shows that bacteria may obtain a large proportion of their genetic diversity through the acquisition of sequences from distantly related organisms, via horizontal gene transfer (HGT) [20, 21, 39, 58, 67, 74, 81, 103]. Furthermore, additional evidence of widespread HGT in plants has emerged recently [5, 6, 73].

Luay Nakhleh
Rice University, e-mail: nakhleh@cs.rice.edu

L.S. Heath and N. Ramakrishnan (eds.), *Problem Solving Handbook in Computational Biology and Bioinformatics*, DOI 10.1007/978-0-387-09760-2_7,

Interspecific recombination is believed to be ubiquitous among viruses [82, 83], and hybrid speciation is a major evolutionary mechanism in plants and groups of fish and frogs [23, 60, 65, 66, 80, 85, 86].

These processes are collectively referred to as *reticulate* evolutionary events and occur at different evolutionary scales: the individual, the population, and the species.

1. Reticulation Between Chromosome Pairs: Meiotic Recombination. During each round of sexual reproduction, the total number of chromosomes must be halved to produce the gametes. The process is called *meiosis*, and during one phase of it the chromosome pairs (sister chromatids) exchange pieces in a precise fashion known as *meiotic recombination*. The net result is chromatids that have two or more evolutionary histories on them. Blocks of chromosomes that share a single evolutionary history are referred to as *haplotype blocks*.

2. Reticulation Within a Lineage: Sexual Recombination. For sexually reproducing organisms, there is recombination of nuclear genomes during each bout of reproduction. Each parent contributes half of its original nuclear genome—one sister chromatid from each chromosome—and each of these chromosomes have themselves undergone meiotic recombination during the process of producing the haploid gametes (sex cells). Because different parts of each parent's contribution to the genome of the next generation may have a different evolutionary history from that of the other parent's contribution, sexual recombination is a form of population-level reticulation. Organellar genomes (mitochondria and chloroplasts) are usually inherited uniparentally so they do not usually undergo any sort of sexual recombination.

3. Reticulation Among Lineages: Horizontal Gene Transfer and Hybrid Speciation. In horizontal (also called lateral) gene transfer (HGT for short), genetic material is transferred from one lineage to another. In an evolutionary scenario involving horizontal transfer, certain sites (specified by a specific substring within the DNA sequence of the species into which the horizontally transferred DNA was inserted) are inherited through horizontal transfer from another species, while all others are inherited from the parent.

In hybrid speciation, which is a form of horizontal transfer, two lineages recombine to create a new species. The new species may have the same number of chromosomes as its parent (*diploid hybridization*) or the sum of the numbers of chromosomes of its parents (*polyploid hybridization*). In a diploid hybridization event, the hybrid inherits one of the two homologs for each chromosome from each of its two parents. Since homologs assort at random into the gametes (sex cells), each has an equal probability of ending up in the hybrid. In polyploid hybridization, both homologs from both parents are contributed to the hybrid. Prior to the hybridization event, each site on the homolog has evolved in a tree-like fashion, although due to meiotic recombination (exchanges between the parental homologs during production of the gametes), different strings of sites may have different histories. Thus, each site in the homologs of the parents of the hybrid evolved in a tree-like fashion on one of the trees contained inside the network representing the hybridization event.

Looking through a macroevolutionary lens (evolution among lineages), only reticulate events at the species level fail to be modeled by a tree. However, looking through a microevolutionary lens (evolution within a lineage), sexual and meiotic recombination fail to be modeled by a bifurcating tree. Since phylogenies are usually constructed at either the population or the species level, meiotic recombination does not cause a species-level reticulate evolutionary history, but it can confound species-level inference of reticulation by producing patterns that have the appearance of species-level reticulation (more on this in Section 4).

In effect, when reticulation occurs, two or more independent evolutionary lineages are combined at some level of biological organization, thus resulting in complex evolutionary relationships that cannot be adequately modeled with trees; instead, *phylogenetic networks* become the appropriate model. Phylogenetic networks are a special class of graphs that allows for multiple paths between pairs of taxa in the phylogeny, and as such provide an extension of phylogenetic trees, in which a unique path exists between any two taxa. Phylogenetic networks come in various flavors, and a variety of methods for reconstructing them have been designed recently. There have been several recent detailed surveys of phylogenetic reconstruction methods [32, 47, 48, 61, 64, 72], some of which identify their similarities and differences. Further, Gambette has created an excellent online resource for documenting all work related to phylogenetic networks [31].

In this chapter, we focus on a specific type of phylogenetic networks, namely *evolutionary phylogenetic networks*, which explicitly model reticulate evolutionary events. Rather than surveying tools and implementations, in this chapter we address issues that are central to accurate detection of reticulate evolution and reconstruction of phylogenetic networks. The rest of this chapter is organized as follows. In Section 2, we define evolutionary phylogenetic networks, discuss their relationships with trees, and outline the general approach for their reconstruction from gene trees. In Section 3, we discuss extensions of three popular optimization criteria, maximum parsimony (MP), maximum compatibility, and maximum likelihood (ML), to the domain of phylogenetic networks. In Section 4, we address various processes that result in patterns that resemble those resulting from reticulate evolutionary events and the need for a framework to distinguish among those processes as a prerequisite to accurate reconstruction of phylogenetic networks. In Section 5, we provide a set of exercises for the reader to gain more understanding of the issues surrounding phylogenetic networks. We conclude in Section 6 with a list of further reading materials that provide in-depth details about other aspects of phylogenetic networks.

2 Phylogenetic Networks and the Trees Within

In this work, we focus on *evolutionary* phylogenetic networks, i.e., networks that model reticulate evolutionary events explicitly. An important assumption underlying all results in this section as well as Section 3 is that the sole cause of gene tree incongruence is reticulate evolution and that a phylogenetic network reconciles gene trees by explicitly modeling reticulate evolutionary events while ignoring discord processes such as lineage sorting. We discuss the implications of incorporating lineage sorting into the framework in Section 4. While much of the literature is on unrooted, undirected networks (and trees), we focus exclusively in this chapter on rooted networks (and trees).

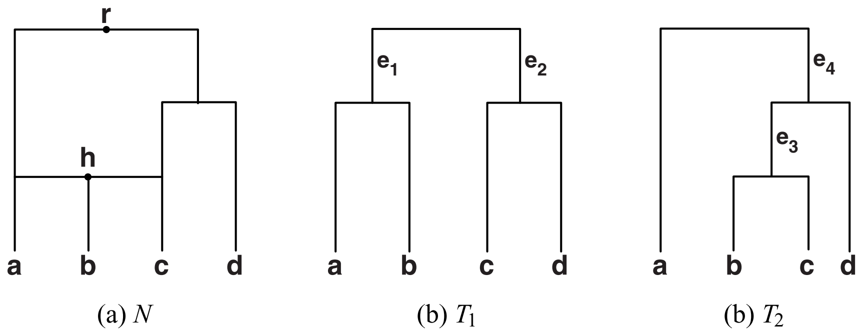
Definition 0.1. A *phylogenetic $\mathscr{X}$-network*, or $\mathscr{X}$-network for short, N is an ordered pair (G,f), where

- $G=(V,E)$ is a directed, acyclic graph (DAG) with $V=\{r\}\cup V_L\cup V_T\cup V_N$, where
  - $indeg(r)=0$ (r is the *root* of N);
  - $\forall v\in V_L$, $indeg(v)=1$ and $outdeg(v)=0$ (V_L are the *leaves* of N);
  - $\forall v\in V_T$, $indeg(v)=1$ and $outdeg(v)\geq 2$ (V_T are the *tree-nodes* of N); and,
  - $\forall v\in V_N$, $indeg(v)=2$ and $outdeg(v)=1$ (V_N are the *network-nodes* of N),

  and $E\subseteq V\times V$ are the network's edges (we distinguish between *network-edges*, edges whose heads are network-nodes, and *tree-edges*, edges whose heads are tree-nodes.
- $f:V_L\rightarrow\mathscr{X}$ is the *leaf-labeling* function, which is a bijection from V_L to $\mathscr{X}$.

Figure 1(a) shows an example of a phylogenetic $\mathscr{X}$-network. Clearly, Definition 0.1 generalizes that of a phylogenetic $\mathscr{X}$-tree; an $\mathscr{X}$-tree is a phylogenetic network with $V_N=\emptyset$.

The semantics of network-nodes are context dependent. For example, in phylogenetics, a network-node may represent a hybrid speciation event, whereas in evolutionary population genetics it may represent a recombination event. While Definition 0.1 requires a network-node to have two parents and a single child, this definition may be relaxed so as to allow for three or more (graph-theoretic) parents and two or more (graph-theoretic) children, which correspond to in- and out-polytomies, respectively; e.g., see the discussions in [70, 76].



(a) N (b) T_1 (b) T_2

Fig. 1 (a) A phylogenetic $\mathscr{X}$-network, rooted at node r, with a single network-node, h, and with $\mathscr{X}=\{a,b,c,d\}$. The trees T_1 (b) and T_2 (c) are the elements of $\mathscr{T}(N)$.

A fundamental observation that underlies reticulate evolution is that the network modeling the evolutionary history of a set of organisms *contains*[1] a set of trees, each of which models the evolutionary histories of certain genomic regions of those organisms. At the lowest level of "atomicity," each nucleotide in the genomes of these organisms has evolved down exactly one of those trees[2]. *The descent of a single nucleotide in a set of organisms could not have followed two different evolutionary histories from the most recent common ancestor of those organisms.* Formally, we can define the set of $\mathscr{X}$-trees contained inside a phylogenetic $\mathscr{X}$-network. Procedure **Induce** in Figure 2 yields one of the trees contained inside a phylogenetic $\mathscr{X}$-network. All trees that can be obtained by applying this procedure to a given phylogenetic $\mathscr{X}$-network N is denoted by $\mathscr{T}(N)$. For the $\mathscr{X}$-network N in Figure 1(a), the set $\mathscr{T}(N) = \{T_1, T_2\}$, where T_1 and T_2 are the two trees shown in Figure 1(b) and 1(c), respectively. Notice that $|\mathscr{T}(N)| = O(b^\ell)$, where b is the maximum indegree of a node in N, and ℓ is the number of network-nodes in N. A tighter bound can be obtained as

```
Induce(N)
Input: Phylogenetic $\mathscr{X}$-network $N = (G, f)$, where $G = (V, E)$.
Output: Phylogenetic $\mathscr{X}$-tree $T = (G', f)$.

1. For each node $v \in V_N$, remove all but one of the edges incoming into $v$;
   let $T = (G', f)$, where $G' = (V', E')$, be the resulting tree.
2. While $\exists u \in V'$ such that $indeg(u) = outdeg(u) = 1$
    a. Let $u$ be such a node with $\{(p,u),(u,c)\} \subseteq E'$;
    b. $V' = V' - \{u\}$;   (* remove a node of indegree and outdegree 1 *)
    c. $E' = E' - \{(p,u),(u,c)\}$;   (* remove its incident edges *)
    d. $E' = E' \cup \{(p,c)\}$;   (* connect its parent to its child *)
3. Return $T$;
```

Fig. 2 Procedure **Induce** for computing a tree in $\mathscr{T}(N)$ for a given phylogenetic $\mathscr{X}$-network N. Observe the random choice of an incoming edge to keep in Step 1. This procedure can be iterated in a deterministic fashion to produce all trees in $\mathscr{T}(N)$ or run non-deterministically a certain number of times to sample from the trees in $\mathscr{T}(N)$.

$$|\mathscr{T}(N)| \leq \prod_{u \in V_N} (indeg(u)). \tag{1}$$

Given an $\mathscr{X}$-network N and an $\mathscr{X}$-tree T, the problem of deciding whether $T \in \mathscr{T}(N)$ is NP-complete [53].

[1] In this context, the term *contain* has been used in the literature interchangeably with two other terms: *induce* and *display*.

[2] Some argue that a forest, rather than a tree, may be a more appropriate model at this atomic level, to allow for events such as insertions and deletions.

Notice that both the **Induce** procedure and the result on the cardinality of $\mathscr{T}(N)$ do not apply when events such as lineage sorting occur; we discuss this in more detail in Section 4.

While computing the set $\mathscr{T}(N)$ for a given $\mathscr{X}$-network is straightforward, computing an $\mathscr{X}$-network N from a set $\mathscr{T}$ of trees is not as straightforward. In fact, this problem is the holy grail of reticulate evolution. First, observe that for a given set $\mathscr{T}$ of $\mathscr{X}$-trees, there may not exist an $\mathscr{X}$-network N such that $\mathscr{T} = \mathscr{T}(N)$ (see Exercise 1); in this case, it is desirable to find an $\mathscr{X}$-network N such that $\mathscr{T} \subseteq \mathscr{T}(N)$. A trivial way to obtain such a network $N = (G, f)$, where $G = (V, E)$, is as follows:

1. $V = \{v_x : x(\neq \emptyset) \subseteq \mathscr{X}\}$. In other words, create one node for each non-empty subset of taxa.
2. $E = \{(v_x, v_y) : v_x, v_y \in V,\ y \subset x\}$.

Clearly, N is an $\mathscr{X}$-network[3] and $\mathscr{T} \subseteq \mathscr{T}(N)$. Baroni *et al.* proposed another "direct" method for constructing a phylogenetic network from a collection of trees [3]. However, while the networks obtained by the method of Baroni *et al.* are smaller in size than those obtained by the method described here, both methods result in a gross overestimation of the extent of reticulation in the evolutionary history.

These observations have been the basis for much work on phylogenetic networks, particularly those with explicit evolutionary implications. In the case of reconstructing *ancestral recombination graphs* (ARGs), the problem has been investigated from the perspective of reconciling the "evolutionary trees" that model the evolution of *single nucleotide polymorphisms*, or SNPs. For reconstructing reticulate evolutionary histories of species, single nucleotides clearly do not provide enough information, and the atomic unit used in this context is a gene. Hereafter, we refer to these units, such as SNPs, genes, haplotype blocks, etc., as *markers*, which are, in essence, the observed biological data from which the phylogenetic network is inferred.

Definition 0.2. The Phylogenetic Network Reconstruction (PNR) Problem

> **Input:** A set of *markers*, $\mathscr{M} = \{M_1, M_2, \ldots, M_k\}$, from a set $\mathscr{X}$ of organisms and a criterion $\boldsymbol{\Phi}$.
> **Output:** A phylogenetic $\mathscr{X}$-network N that models the evolution of $\mathscr{M}$ and that is optimal under criterion $\boldsymbol{\Phi}$.

For example, one version of the problem of inferring ancestral recombination graphs (ARGs) can be formulated as an instance of PNR if one takes $\mathscr{M}$ to be the set of SNPs, and $\boldsymbol{\Phi}$ to be the criterion "N contains the minimum number of network-nodes and every SNP is compatible with at least one tree in $\mathscr{T}(N)$." As another example, one version of the problem of inferring species evolutionary networks can be formulated as an instance of PNR if, given a set $W = \{T_1, \ldots, T_k\}$ of trees with T_i being the *gene tree* of gene M_i, the criterion $\boldsymbol{\Phi}$ is taken to be "N contains the minimum number of network-nodes and $T_i \in \mathscr{T}(N)$ for every $T_i \in W$."

[3] This construction does not ensure that the leaves have indegree of 1, which is one of the requirements in Definition 0.1, but the construction can be extended in a straightforward manner to take care of this.

2.1 Combining Trees Into a Network via SPR Operations

One of the most commonly pursued approaches for reconstructing phylogenetic networks is based on reconciling "gene trees," under the assumption that incongruities, or disagreements, among these trees are caused by only reticulate evolutionary events, such as horizontal gene transfer or hybrid speciations. In this case, several methods have been developed for inferring a lower bound on the number of reticulation events by identifying the minimum number of *subtree prune and regraft*, or SPR, operations required to transform one tree into the other. As the name indicates, an SPR operation applied to tree T cuts, or prunes, a subtree t of T, yielding a tree T', and attaches, or regrafts, it from its root to another branch in T' [1]; see Figure 3 for an illustration. The SPR distance between two trees is the minimum number of SPR moves required to transform one tree into the other. For example, the SPR distance between the two trees in Figure 1 is 1, since a single SPR move is required, as illustrated in Figure 3.

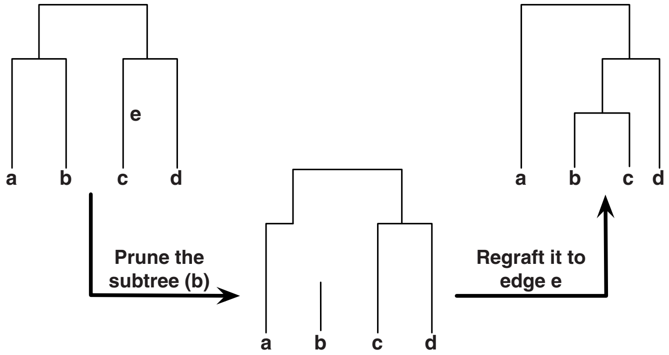


Fig. 3 An illustration of the *subtree prune and regraft*, or SPR, operation. The subtree that contains only the leaf b is pruned from the tree on the left, thus resulting in a forest of two trees, shown in the middle, and then the subtree is regrafted as a sibling of c, resulting in the tree on the right. Observe that the SPR distance between the two trees on the left and right is 1, and both trees can be reconciled in a phylogenetic network with one network-node, as can be seen in Figure 1.

The problem of computing the SPR distance between two rooted trees has been shown to be NP-hard as well as fixed-parameter tractable [7]. Examples of exact algorithms and heuristics for reconciling trees via SPR operations include the exact algorithm of Bordewich and Semple [7], the exact algorithm of Wu [105], HorizTrans [38], RIATA-HGT [78], EEEP [4], HorizStory [62], and the method of Goloboff as implemented in the TNT software package [33]. For the most part, these methods are aimed at finding the phylogenetic network N with the minimum number of network-nodes that contains the pair of input trees. For example, the network N in Figure 1(a) is the only phylogenetic network with a single network-node that contains both trees T_1 and T_2 in Figure 1.

There are several limitations with using the SPR distance as a proxy for the amount of reticulation, as well as with the methods listed above for estimating this distance. We discuss some of those here, and discuss the issue of *time-consistency* of SPR moves in the next section.

It is worth mentioning that methods that attempt to find minimal sets of SPR moves to reconcile a pair of trees are in fact attempts at approximating the true number of reticulation events in the evolutionary history. However, while the SPR distance provides a lower bound on this number, recent results have shown that the SPR distance can provide a value that is arbitrarily smaller than the true amount of reticulation [2, 46].

The tools listed above all assume $k = 2$ (i.e., they solve the problem for a pair of trees) and assume that each of the two trees has exactly $|\mathscr{X}|$ leaves, each labeled uniquely by one label from $\mathscr{X}$. In other words, these tools do not solve the problem, in terms of computing a minimal network, for more than two trees, nor do they allow for trees with different leaf-sets. Both of these present practical limitations to the use of the methods in practice, particularly the latter, since, in general, there is no guarantee that a 1-1 correspondence exists between the leaves of the (species) phylogenetic network and those of the gene trees.

A very important issue that tools for combining trees into a network must account for is the potential multiplicity of different, optimal (minimal, in this case) networks. Than *et al.* [100] showed that the number of minimal networks that reconcile a given pair of trees may be exponential in the minimum number of reticulation events required.

Last but not least, reconstructed gene trees are often non-binary (which mostly indicates soft polytomies[4]). The reconciliation problem becomes more complicated when non-binary trees are concerned. In this case, one objective is to *simultaneously* resolve the trees and infer the minimum number of reticulation events. The number of resolutions of non-binary tree is exponential in the degree of the nodes, and hence efficient techniques are required for solving this problem. Than and Nakhleh [99] provided a heuristic for solving several cases of this problem, which are implemented in the PhyloNet package [101] as an extension of the RIATA-HGT method [78].

2.2 Totally-ordered Trees and Time-consistent SPR Operations

In our discussion thus far of the SPR operation and its induced distance, we have considered only the topologies of a pair of trees. However, when times at the internal nodes of the species and gene trees are known (in the former case, those times would indicate the divergence time of the species from their common ancestors, and in the latter case those times would indicate the times of the coalescence events),

[4] In a rooted phylogenetic tree, a polytomy is a node with more than two children. There are two types of polytomies: a *hard polytomy* indicates the hypothesis that the speciation event gave rise to multiple lineages, whereas a *soft polytomy* indicates the lack of knowledge to resolve a multifurcating node into a sequence of bifurcating nodes.

the situation becomes more complicated. Rooted trees in which internal nodes are totally ordered are called *ordered tree* [90].

When ordered trees are considered, two crucial issues arise:

1. Topologically identical or similar trees may be very different when branch lengths are considered (S. Edwards recently labeled such phenomenon "branch length heterogeneity" [22], though in the different context of lineage sorting), and
2. certain SPR moves may not be *time consistent*.

We elaborate on these two issues in a few examples. Consider the two trees in Figure 4. Topologically, the two trees are identical. However, considering the trees on the left and right to be the species and gene trees, respectively, the species a and b diverged at time T_2 (similarly for species c and d), while their genes coalesced at time T_1, which is different from T_2. This is a scenario of branch length heterogeneity, and the trees, when viewed as ordered trees, are different.

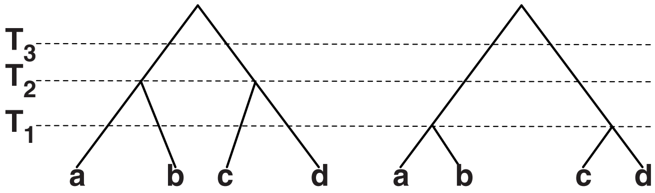


Fig. 4 Two phylogenetic trees that require no SPR moves to transform into each other, when only the topologies are considered. However, when times at internal nodes are considered, the two trees are different, and require a minimum number of two SPR moves, as shown in Figure 5. The horizontal dashed lines represent times.

In this case, the true SPR distance is not zero, but rather two, as illustrated in Figure 5.

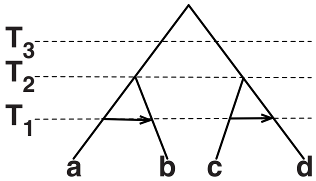


Fig. 5 Two SPR moves required to transform the tree on the left in Figure 4 into the one on the right, when times at internal nodes are taken into account. The horizontal dashed lines represent times.

For the second issue, consider the species and gene trees shown in Figure 6 (left and right, respectively). When their topologies are compared, a single SPR move suffices to transform the species tree into the gene tree, as shown in Figure 7. However, notice that in this scenario, the transfer of the genetic material took place between two organisms that do not co-exist in time. In other words, this SPR move is not time consistent.

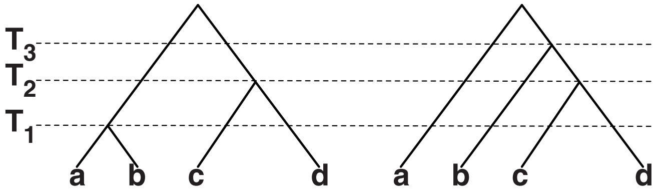


Fig. 6 Two trees that differ in the placement of b, thus requiring a single SPR move to transform the tree on the left into the one on the right, as shown in Figure 7(a), when only the topologies of the trees are considered. However, such a move is not time-consistent since the "donor" (tail of the HGT edge) and "recipient" (head of the HGT edge) do not co-exist in time. The horizontal dashed lines represent times.

An important question in this case is whether such an SPR move should be ruled out in a species/gene tree reconciliation scenario. While the scenario, as drawn in Figure 7(a), contains a time inconsistent SPR move, this inconsistency may be explained as an artifact of *incomplete taxon sampling*, as we now illustrate.

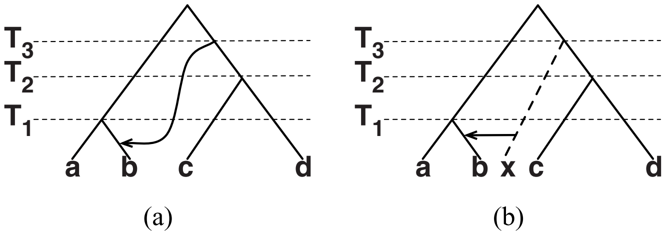


Fig. 7 (a) The SPR move required for the scenario in Figure 6. This SPR move is not time-consistent. (b) The SPR move can be viewed as time-consistent if incomplete taxon sampling occurs; in this case, the horizontal transfer occurred from a taxon x, which is contemporaneous with b yet was not sampled when the the species and gene trees were reconstructed. The horizontal dashed lines represent times.

Consider the case in which the horizontal transfer occurred from species x to species b, where x was a sibling of the clade (c,d) yet was not sampled (or became extinct after the horizontal transfer event) in the evolutionary analysis. This case is shown in Figure 7(b). In this scenario, while the SPR moves necessary to transform the species tree into the gene tree is seemingly time inconsistent, it is in fact a reflection of incomplete taxon sampling, or even a true biological hypothesis—that of the extinction of species x. Determining whether a time inconsistent SPR move is truly so or is merely a reflection of incomplete taxon sampling (or extinction) is a very challenging question.

It is important to note, though, that not all time inconsistent SPR moves can be justified with the incomplete taxon sampling scenario. Consider the species and gene trees in Figure 8 (left and right, respectively). In this case, a single SPR move, pruning the clade (b,c) and regrafting it as a sibling of d, would reconcile the two trees, as shown in Figure 9(a). Clearly, this SPR move is time inconsistent. Unlike the previous case, incomplete taxon sampling cannot explain the inconsistency in this scenario, since no matter how we augment the species tree with "phantoms" of missing taxa, the source and destination of the SPR move cannot be made contemporaneous. Instead, a scenario involving two time consistent SPR moves may be the correct one, as illustrated in Figure 9(b).

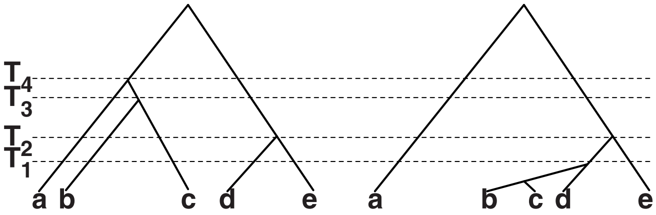
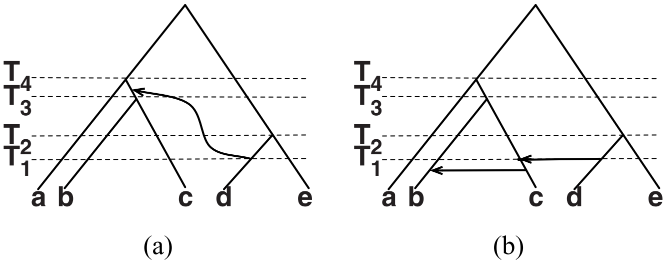


Fig. 8 Two trees that differ in the placement of clade (b,c), thus requiring a single SPR move to transform the tree on the left into the one on the right, as shown in Figure 9(a), when only the topologies of the trees are considered. However, such a move is not time-consistent since the donor and recipient do not co-exist in time. The horizontal dashed lines represent times.



(a) (b)

Fig. 9 (a) The SPR move required for the scenario in Figure 8. This SPR move is not time-consistent. (b) A solution of two time-consistent SPR moves that explains the scenario in Figure 8. The horizontal dashed lines represent times.

3 Optimization Criteria for Inferring and Evaluating Phylogenetic Networks

The relationship between a phylogenetic network N and its constituent trees, $\mathscr{T}(N)$, allows for extending sequence-based optimization criteria from phylogenetic trees to phylogenetic networks. Such extensions are based on the fact that, at the lowest level of atomicity in genetic inheritance, a nucleotide in the genomes of a set of species evolves down a single tree, even if the evolutionary history of the species is best modeled by a network[5]. This, in essence, is the concept of *positional homology*. In this section, we discuss extensions to the maximum parsimony, maximum compatibility, and maximum likelihood criteria.

Let T be an $\mathscr{X}$-tree with leaf-set $\mathscr{L}$, and let Σ be an alphabet (e.g., $\Sigma = \{A,C,T,G\}$ for DNA). A function $\lambda : \mathscr{L} \to \Sigma$ is called a *state assignment function* for tree T over alphabet Σ. The function $\hat{\lambda} : V(T) \to \Sigma$ is an extension of λ on T if it agrees with λ on the leaves of T (i.e., if $\hat{\lambda}(v) = \lambda(v)$ for every $v \in \mathscr{L}$). In a similar way, we define a function $\lambda^k : \mathscr{L} \to \Sigma^k$ and an extension $\hat{\lambda}^k : V(T) \to \Sigma^k$. The latter function is called a *labeling* of T, and it denotes the labeling of all nodes of a tree T with sequences of length k over alphabet Σ. Given a labeling $\hat{\lambda}^k$, we denote by $d_e(\hat{\lambda}^k)$ the Hamming distance (or any edit distance) between the two sequences labeling the two endpoints of edge $e \in E(T)$. We define the state assignment and labeling functions for an $\mathscr{X}$-network similarly. The difference between the labeling of a tree and that of a network lies in the interpretation of sequence evolution. Let (u,v) be an edge in a phylogenetic tree with $x = \hat{\lambda}^k(u)$ and $y = \hat{\lambda}^k(v)$. Then, the state at position i in sequence y is the result of zero or more mutations on the state at position i in sequence x. In a phylogenetic network, this interpretation is slightly more involved. Assume edge (u,v) in a phylogenetic network, with x and y defined as before. If $indeg(v) = 1$, then the relationship between the states at position i in sequences x and y is identical to that in trees. However, if $indeg(v) = m$, where $m > 1$, then the state at position i in sequence y is the result of zero or more mutations on the state at position i in exactly one of the sequences labeling the m parents of v. This labeling and interpretation serve as the basis for extending sequence-based optimization criteria from trees to networks.

3.1 Maximum Parsimony of Phylogenetic Networks

Roughly speaking, the maximum parsimony criterion is a reflection of Occam's razor; that is, the best solution is the simplest. In the context of phylogenetics, the maximum parsimony criterion seeks the tree on a given set of genomic sequences such that the tree minimizes the overall number of mutations along all edges of the tree. This is formalized as follows.

[5] The same comment in Footnote 2 applies here.

Definition 0.3. The parsimony length of a phylogenetic tree T with a labeling λ^k is $PS(T,\lambda^k) = \min_{\hat{\lambda}^k \in \hat{\Lambda}^k} [\sum_{e \in E(T)} d_e(\hat{\lambda}^k)]$, where $\hat{\Lambda}^k$ is the set of all possible extensions of λ^k.

We denote by $PS_i(T,\lambda^k)$ the parsimony length of tree T with respect to site i. Given a labeling λ^k of a set $\mathscr{X}$ of taxa, the maximum parsimony (MP) problem for phylogenetic trees amounts to solving

$$T^* = \text{argmin}_T PS(T,\lambda^k), \tag{2}$$

where T ranges over all $\mathscr{X}$-trees. There is a polynomial time algorithm for computing the parsimony length of a fixed $\mathscr{X}$-tree [29], while solving the MP problem in general is NP-hard [18, 30].

In the early 1990's, Jotun Hein introduced an extension of the maximum parsimony (MP) criterion to model the evolutionary history of a set of sequences in the presence of recombination [40, 41]. Recently, Nakhleh and colleagues gave a mathematical formulation of the MP criterion for phylogenetic networks and devised computationally efficient solutions aimed at reconstructing and evaluating the quality of phylogenetic networks under the MP criterion [49, 51, 52]. The parsimony length of a phylogenetic network with respect to a set of sequences is defined as follows.

Definition 0.4. The parsimony length of a phylogenetic network N with a labeling λ^k of the leaves of N is

$$PS(N,\lambda^k) = \sum_{1 \le i \le k} \left[\min_{T \in \mathscr{T}(N)} PS_i(T,\lambda^k) \right].$$

Notice that this definition of the parsimony length allows for the rather biologically unrealistic scenario of switching back and forth between different trees for consecutive sites. For example, for $k = 10$, the definition may lead to the scenario in which sites 1, 3, 5, 7, and 9 are best fit by tree T' and sites 2, 4, 6, 8, and 10 are best fit by tree T'', for two different trees T' and T''. This was addressed in practice in the sequence of papers by Jin *et al.* by doing the computation on a block-by-block, rather than site-by-site, basis. Another way to address this issue is to introduce a penalty for switching among trees. As the parsimony criterion is based on the assumption of rare events (e.g., [26, 27]), a reticulation event may be best modeled as causing a penalty of one change (J. Felsenstein, personal communication).

Given a labeling λ^k of a set $\mathscr{X}$ of taxa, the maximum parsimony (MP) problem for phylogenetic networks amounts to solving

$$N^* = argmin_N PS(N,\lambda^k), \tag{3}$$

where N ranges over all $\mathscr{X}$-networks. Unlike the case of trees, the problem of computing the parsimony length of a fixed $\mathscr{X}$-network is NP-hard [49], and the problem of solving the MP problem for phylogenetic networks is NP-hard as well, as it contains the MP problem for trees as a special case.

Let N be an $\mathscr{X}$-network, and let N' be another $\mathscr{X}$-network obtained by adding a set H of edges to N, where each edge in H is posited between a pair of edges whose heads are tree-nodes in N. Then, we have

$$\mathscr{T}(N) \subseteq \mathscr{T}(N').$$

This result is illustrated in Figure 10.

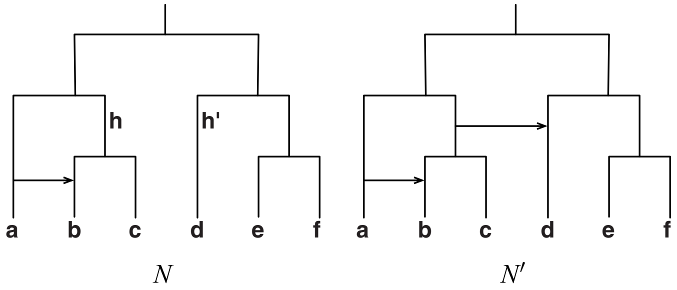


Fig. 10 Two $\mathscr{X}$-networks N and N' such that N' is obtained by adding an additional edge to N from edge h to edge h'. We have $\mathscr{T}(N) = \{T_1, T_2\}$ and $\mathscr{T}(N') = \{T_1, T_2, T_3, T_4\}$, where $T_1 = ((a,(b,c)),(d,(e,f)))$, $T_2 = (((a,b),c),(d,(e,f)))$, $T_3 = ((a,((b,c),d)),(e,f))$, and $T_4 = (((a,b),(c,d)),(e,f))$. Clearly, $\mathscr{T}(N) \subseteq \mathscr{T}(N')$.

From this fact it follows that, for a given labeling λ^k of a set $\mathscr{X}$ of taxa, we have

$$PS(N', \lambda^k) \leq PS(N, \lambda^k).$$

This simple observation has a significant implication on the use of the MP criterion for inferring networks, as defined above. It basically implies that adding more edges to a network "never hurts" under the MP criterion as defined above: the parsimony length either decreases or stays the same as more edges are added. This in turn implies that while making networks more "complex" improves their parsimony lengths, using the MP criterion in this fashion would inevitably result in a gross over-estimation of the amount of reticulation in the evolutionary history of a data set. This had led to refining the definition of the MP criterion so that adding edges to a network is accepted only if the parsimony length is improved beyond a given threshold [51]. Currently, such a threshold is dataset-specific and is determined by inspection of the trend of parsimony length decrease as the complexity of networks is increased. Such an approach has produced very promising results, on both synthetic and biological data sets [51, 98].

3.2 Character Compatibility of Phylogenetic Networks

Two models of sequence evolution that have been central in population genetics, and which have been assumed to underlie a special type of phylogenetic networks are the *infinite-allele model* and *infinite-site model* . The infinite-allele model, proposed by Kimura and Crow [56], assumes that each mutation at a site results in a state that is different from any preexisting state at that site in the population. The infinite-site model, proposed by Kimura [55], assumes that the sequences are very long and that the mutation rate per site is low so that each site mutates at most once. These two models can be formulated within the parsimony framework. If a site i evolves down a tree T under the infinite-allele model, and m distinct states are observed at site i in the leaves of T, then the parsimony length of T with respect to site i is $m-1$. If site i evolves under the infinite-site model, then the parsimony length of T with respect to site i is either 0 (no mutations occurred at site i) or 1 (exactly one mutation occurred). In the phylogenetics jargon, a site that evolves down a tree T under either infinite-allele or infinite site model is said to be *compatible* with the tree T. A tree T for which all sites in the sequences labeling its leaves are compatible is called a *perfect phylogeny*. Gusfield provided an $O(nm)$ algorithm for determining whether there exists a perfect phylogeny for a set of n binary sequences, each of length m, and reconstructing such a perfect phylogeny if it exists [35], thus improving on an earlier $O(nm^2)$ algorithm [24, 68].

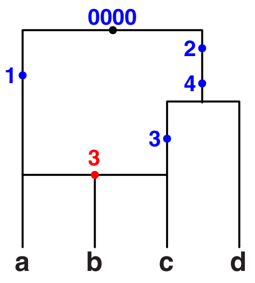
Barring any (meiotic) *recombination* events, the evolutionary history of a sequence of sites under the infinite-site model is modeled by a tree. However, when recombination occurs, the evolutionary histories of sites to the left and right of a recombination breakpoint follow different paths in their ancestries, thus giving rise to a phylogenetic network model. The compatibility criterion can be extended to phylogenetic networks in a fashion similar to that of extending the MP criterion. We say that a site is compatible with a phylogenetic network N if it is compatible with at least one of the trees in $\mathscr{T}(N)$. Determining if a site is compatible with a phylogenetic network is NP-Complete [53]. An *ancestral recombination graph* [34], or ARG for short, is a phylogenetic network that models the evolution of a set of sequences under the infinite-site model, in which:

- each edge is labeled by a set of numbers denoting the sites that mutate along that edge,
- each node of indegree 2 is labeled by a number denoting the recombination breakpoint giving rise to that network-node, and
- each site in the sequences is compatible with the network.

Figure 11 shows an ARG modeling the evolutionary history of a set of four sequences under the infinite-site model. ARGs have also been referred to as *perfect phylogenetic networks* [102]. Much work has been done on reconstructing *minimal* ARGs, i.e., ARGs with the minimum number of nodes of indegree 2 to model the evolution of a set of binary sequences under the infinite-site model; e.g., see [36, 37, 91, 92, 93, 94]. Recently, Willson provided a new method for

	c_1	c_2	c_3	c_4
a	1	0	0	0
b	1	0	1	1
c	0	1	1	1
d	0	1	0	1

(a)



(b)

Fig. 11 (a) A data set of four binary sequences **a**, **b**, **c**, and **d**. (b) An ARG showing the evolutionary history of the four sequences from the ancestral sequence 0000 under the infinite-site model. The solid circle at the node of indegree 2 indicates a recombination event, and the value 3 indicates that the states of sites 1 and 2 (which are 1 and 0, respectively) were inherited from the left parent, whereas the states of sites 3 and 4 (which are 1 and 1, respectively) were inherited from the right parent, thus forming the sequence 1011 at **b**. The other solid circles indicate mutations, where the numbers associated with them indicate the site at which each mutation occurred.

reconstructing certain phylogenetic networks from binary sequences when *back-mutations* are allowed to occur at network-nodes [104].

While we focused on binary characters in the preceding discussion, perfect phylogenetic networks can be defined on multi-state characters as well. Let $\lambda : \mathscr{L} \rightarrow \Sigma$ be a leaf-labeling of a tree T, with $\Sigma' \subseteq \Sigma$ being the character states that are observed at the leaves of T (not all character states in Σ may be observed at the leaves, and hence the need for Σ'). We say that λ is compatible on T if there exists an extension $\hat{\lambda}$ such that

$$\sum_{e \in E(T)} d_e(\hat{\lambda}) = |\Sigma'| - 1.$$

We say that λ is compatible with a phylogenetic network N if it is compatible with at least one of the trees in $\mathscr{T}(N)$. Character compatibility on a tree and on a network can be extended in a straightforward manner to sequences of characters (λ^k). Figure 12(a) shows a tree whose leaves are labeled by sequences of length 2 over the alphabet $\Sigma = \{1,2,3,4\}$. For the first character (site), we have $\Sigma'_1 = \{1,2,3\}$ and for the second we have $\Sigma'_2 = \{2,3,4\}$. The first character is compatible with the tree, whereas the second is not. When a single reticulation event is added to the tree, as shown in Figure 12(b), we obtain a perfect phylogenetic network for the sequences labeling the leaves; see Exercise 5.

Nakhleh *et al.* proposed multi-state perfect phylogenetic networks[6] to model the evolutionary histories of natural languages in the presence of borrowing [77]. The *Character Compatibility on Phylogenetic Networks* Problem is to decide whether a given phylogenetic network is a perfect phylogenetic network for a set C of characters (alternatively, a leaf-labeling λ^k). This problem has been shown to be NP-

[6] In [77], network-edges were allowed to be bi-directional.

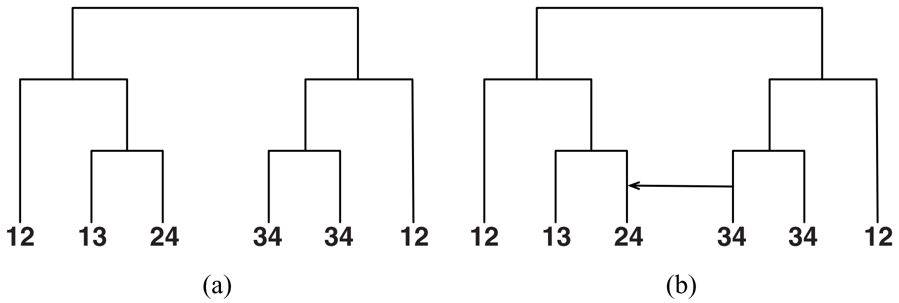


Fig. 12 (a) A phylogenetic tree leaf-labeled by sequences of length 2 over the alphabet $\Sigma = \{1,2,3,4\}$. The first character is compatible on the tree, whereas the second is not. (b) A perfect phylogenetic network obtained from the tree by adding a single reticulation event.

hard [75] even for binary characters. Kanj *et al.* provided an efficient parameterized algorithm for the binary case of this problem [54].

3.2.1 Binary Character Compatibility and Combining Trees into a Network

There is an elegant connection between the problem of combining a set of trees into a network and the problem of inferring a perfect phylogenetic network (with only uni-directional edges) for a set of binary sequences. Let $\mathscr{T} = \{T_1, T_2, \ldots, T_m\}$ be a set of (rooted) $\mathscr{X}$-trees. For each edge e in a tree $T_i \in \mathscr{T}$, define a binary site c_e with its states assigned as follows for each $x \in \mathscr{X}$:

$$c_e(x) = \begin{cases} 1, & x \text{ under } e; \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The collection $C = \cup_{T_i \in \mathscr{T}} \{c_e : e \in E(T_i)\}$ is called the *character encoding* of the trees in $\mathscr{T}$. If the trees in $\mathscr{T}$ contain p distinct edges (two edges are distinct if they define different clusters of leaves), then C contains p distinct sites, and each taxon $x \in \mathscr{X}$ is associated with a binary sequence s_x of length p. The main result here is that if N is a network such that $\mathscr{T} \subseteq \mathscr{T}(N)$ then N is a perfect phylogenetic network for the set C, which is the character encoding of $\mathscr{T}$.

	e_1	e_2	e_3	e_4
a	1	0	0	0
b	1	0	1	1
c	0	1	1	1
d	0	1	0	1

Fig. 13 The character encoding of the two trees in Figure 1(b) and 1(c). The resulting matrix is identical to that of the binary sequences in Figure 11.

Figure 13 shows the character encoding of the two trees in Figure 1. Indeed, the sequences in Figure 13 are compatible with the phylogenetic network in Figure 11, which is identical (in terms of topology) to the phylogenetic network N in Figure 1(a) that contains the two trees.

3.3 Maximum Likelihood of Phylogenetic Networks

Extending the maximum likelihood (ML) criterion to phylogenetic networks is done in a similar fashion to that used in the MP criterion, with the additional details about the probabilistic setting in which to interpret the trees of a network and summarize the likelihood scores computed on these trees.

Assuming independence among sites, the overall likelihood of a set S of aligned sequences, given by the labeling function λ^k, given a tree topology ψ and a model M (branch lengths and model of sequence evolution), is the product of the probability of the labeling of every site i given ψ and M:

$$L(\lambda^k|\psi,M) = \prod_{i=1}^{k} L(\lambda^k[i]|\psi,M), \tag{5}$$

where k is the number of sites, and $L(\lambda^k[i]|\psi,M)$ can be defined in two ways:

- For *(average) likelihood* [95], L_{avg}, we have:

$$\sum_{\hat{\Lambda}^k} \left[\mathbf{P}(root) \cdot \prod_{e \in E(T)} \mathbf{P}_e(t_e) \right], \tag{6}$$

  where $\hat{\Lambda}^k$ is the set of all possible extensions of λ^k, and $\mathbf{P}_e(t_e)$ denotes the probability of observing the sequences at the two endpoints of edge e whose branch length is t_e.
- For *ancestral likelihood* [84], L_{anc}, we have:

$$\max_{\hat{\Lambda}^k} \left[\mathbf{P}(root) \cdot \prod_{e \in E(T)} \mathbf{P}_e(t_e) \right]. \tag{7}$$

Given a labeling λ^k of a set $\mathscr{X}$ of taxa, the maximum likelihood (ML) problem for phylogenetic trees amounts to solving

$$(\psi^*, M^*) = \text{argmax}_{\psi,M} L(\lambda^k|\psi,M), \tag{8}$$

where ψ ranges over all $\mathscr{X}$-tree topologies, and M ranges over all combinations of branch lengths and models of sequence evolution. When all elements of this combination are specified, scoring the likelihood can be done in polynomial time using

Felsenstein's "pruning" algorithm [28]. Solving the ML problem in general is NP-hard [16].

Lathrop defined a maximum likelihood criterion for phylogenetic inference of populations when some of those populations are hybridized (in this context, hybridization corresponds to *admixture*) [59]. Strimmer and Moulton defined the maximum likelihood criterion for *splits networks*, once their edges are oriented so as to produce a rooted, directed, acyclic, graph [96]. Jin *et al.* defined ML criteria for evolutionary phylogenetic networks [50], which we review here.

Let N be an $\mathscr{X}$-network in which network-nodes have indegree 2 (the results can be generalized in a straightforward way to networks with nodes whose indegree is higher than 2), and let $\mathscr{R} = \{p_i = (e^i_l, e^i_r) : e^i_l, e^i_r \in E(N),\ e^i_l = (x,v),\ e^i_r = (y,v)$, and $x \neq y\}$, with $r = |\mathscr{R}|$. In other words, $\mathscr{R}$ is the set of pairs of edges where each pair is incident into the same network node. Further, we associate with each pair $p_i \in \mathscr{R}$ parameter $\gamma_i \in [0,1]$ which denotes the probability of choosing the "left" edge e^i_l (the probability of choosing the "right" edge e^i_r is $(1-\gamma_i)$). These probabilities are to be estimated from the sequence data, and can be interpreted as the proportion of sites (of the sequence at a network-node) inherited from one of the parents [96]. When multiple loci are involved in the analysis, these probabilities can denote the proportion of the genome arising from a particular parent [69]; see Section 4.3. In the case of admixture, these probabilities correspond to the proportion of the population derived from a particular ancestral population [59]. For example, consider the phylogenetic network N in Figure 14. For this network, we have $\mathscr{R} = \{p_1 = ((u,x),(v,x)), p_2 = ((w,y),(z,y))\}$, parameter γ_1 associated with p_1 (which denotes the probability of taking edge (u,x) for certain sites in the sequence at node x), and parameter γ_2 associated with p_2 (which denotes the probability of taking edge (w,y) for certain sites in the sequence at node y).

Let $T \in \mathscr{T}(N)$. A *characteristic set* of tree T is a set φ_T of size r that contains exactly one edge from every pair in $\mathscr{R}$ such that when all network-edges except for those in φ_T are removed from network N in Step 1 of procedure **Induce** in Figure 2, the procedure yields tree T. For the network N and its induced trees shown in Figure 14, we have $\varphi_{T_1} = \{(v,x),(z,y)\}$, $\varphi_{T_2} = \{(u,x),(z,y)\}$, $\varphi_{T_3} = \{(v,x),(w,y)\}$, and $\varphi_{T_4} = \{(u,x),(w,y)\}$.

Notice that multiple characteristic sets may exist for the same tree T; in this case, we denote the set of all characteristic sets by Φ_T. Then, the probability of a tree T, given network N and leaf-labeling λ^k is

$$\mathbf{P}(T|N,\lambda^k) = \sum_{\varphi_T \in \Phi_T} \left[\prod_{e^i_l \in \varphi_T} \gamma_i \prod_{e^j_r \in \varphi_T} (1-\gamma_j) \right]. \qquad (9)$$

In other words, the probability of inducing a tree T by network N is the product of the probabilities of all the network-edges used to induce T. The summation in the formula is to account for cases when there exist multiple ways to induce the tree T. The probabilities of the four trees in Figure 14 are given in the caption of the figure.

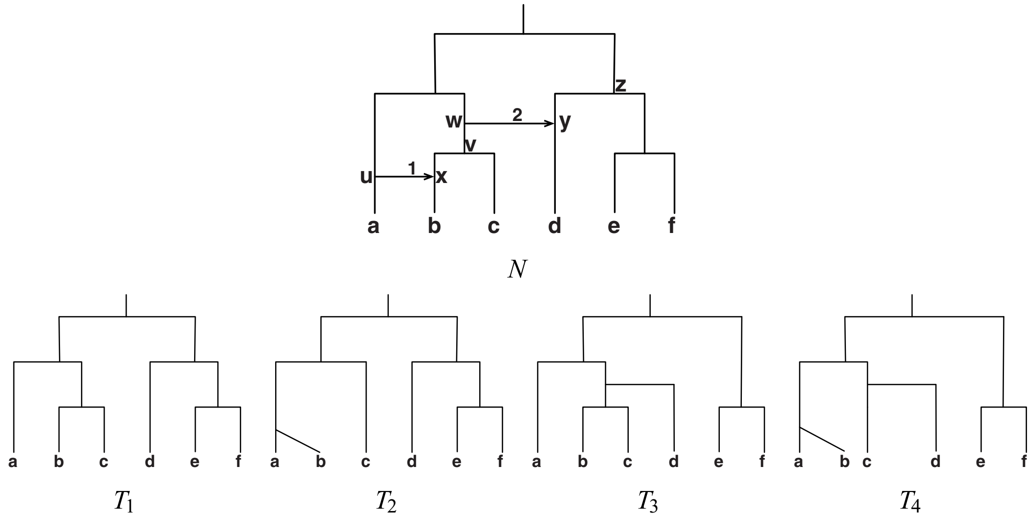


Fig. 14 A phylogenetic network (top) and the four trees it induces (bottom). Using Formula (9), we have $\mathbf{P}(T_1|N) = (1-\gamma_1)(1-\gamma_2)$, $\mathbf{P}(T_2|N) = \gamma_1(1-\gamma_2)$, $\mathbf{P}(T_3|N) = (1-\gamma_1)\gamma_2$, and $\mathbf{P}(T_4|N) = \gamma_1\gamma_2$.

We are now in position to define likelihood criteria for phylogenetic networks. The likelihood of a phylogenetic network with respect to a set of sequences is defined as follows.

Definition 0.5. The likelihood of a phylogenetic network N with a labeling λ^k of the leaves of N is

$$L(\lambda^k|N,M) = \sum_{T \in \mathscr{T}(N)} \left[\mathbf{P}(T|N,\lambda^k) \cdot L(\lambda^k|T,M_T)\right], \tag{10}$$

where M is the model (branch lengths, probabilities γ_i, and model of sequence evolution), M_T is the "restriction" of M to tree T, and $L(\lambda^k|N,M)$ can be either the average or ancestral likelihood functions.

Given a labeling λ^k of a set $\mathscr{X}$ of taxa, the maximum likelihood problem for phylogenetic networks can be defined so as to solve

$$(N^*,M^*) = \text{argmax}_{N,M} L(\lambda^k|N,M), \tag{11}$$

where N ranges over all $\mathscr{X}$-network topologies, and M ranges over all combinations of branch lengths, probabilities γ_i, and models of sequence evolution.

Notice that, while the likelihood of a network, as given by Definition 0.5, is an average of the likelihood of all trees within the networks, we can modify this definition so that the likelihood of a network is the best over all trees, which is analogous to the way we defined the parsimony length of a network above. In this case, we have

$$L(\lambda^k|N,M) = \max_{T \in \mathscr{T}(N)} \left[\mathbf{P}(T|N,\lambda^k) \cdot L(\lambda^k|T,M_T)\right].$$

This definition would be more appropriate for inferring ancestral states on a phylogenetic network.

Finally, the type of input data further refines the versions of the ML problems, as outlined in [50]. This results in several formulations of ML criteria for phylogenetic networks, where these formulations amount to the combinations of tree likelihood type (ancestral vs. average), tree selection criterion (average vs. maximum), and input data.

Problem 0.1. (The Tiny ML Problem)

Input: The full model M of an $\mathscr{X}$-network N, and a labeling λ^k of the leaves.
Output: The labeling $\hat{\lambda}^k$ that maximizes the likelihood of the network.

Problem 0.2. (The Small ML Problem)

Input: The topology of a phylogenetic network N and a labeling λ^k of the leaves.
Output: The branch lengths, edge probabilities, and labeling $\hat{\lambda}^k$ that maximize the likelihood of the network.

Problem 0.3. (The Big ML Problem)

Input: The labeling λ^k of a set $\mathscr{X}$ of taxa.
Output: A full model M of an $\mathscr{X}$-network N that maximizes $L(\lambda^k|N,M)$.

4 To Network, or Not to Network, That Is the Question

In our discussion thus far, we have made an important assumption: incongruities and incompatibilities in the data are due to reticulate evolutionary events and therefore should be reconciled by using a phylogenetic network. We assumed that gene trees disagree due to the occurrence of events such as horizontal gene transfer, and sought a network that reconciles them. In the case of ancestral recombination graphs and perfect phylogenetic networks, we assumed that if a perfect phylogenetic tree does not exist for a set of sequences, then that is an indication of the occurrence of intralocus recombination [45], and hence a network, rather than a tree, is sought as a model of the evolutionary history. However, this assumption must be inspected carefully and thoroughly before phylogenetic network reconstruction is attempted. Several ways exist for explaining the evolution of a data set without invoking reticulate evolutionary events:

- In the analysis of biological data, gene trees are unknown and reconstructed from sequence data. These reconstructions of the trees may have errors in them, in the form of wrong edges. When compared to a species tree, these wrong edges masquerade as true incongruities, triggering the false inference of reticulate evolutionary events, and sometimes they may in fact hide true incongruities, thus resulting in an underestimation of the amount of reticulation in the data; e.g., see [100].

- As Figure 15(a) shows, a gene tree may disagree with a species tree due to a combination of duplication and loss events that took place during the evolution of the gene. In this case, and notwithstanding the incongruities among gene trees, these trees need be reconciled into a tree, not a network.

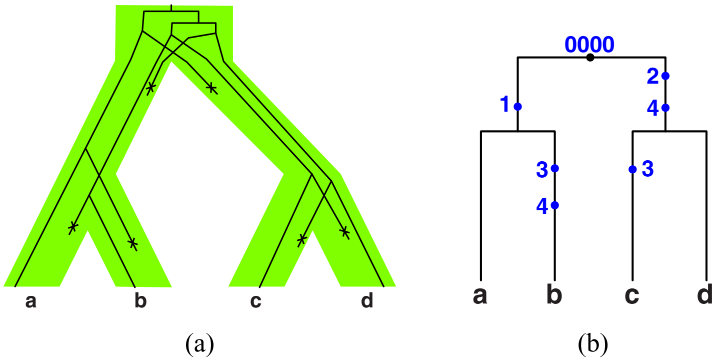


Fig. 15 (a) A gene tree (solid lines) evolving within the branches of the species tree, where the gene tree topology is identical to that of T_2 in Figure 1(b). The gene tree differs from the species tree due to multiple gene duplication and loss events. (b) A phylogenetic tree that models the evolutionary history of the sequences in Figure 11(a) from the ancestral sequence 0000, while violating the infinite-site model assumptions. In this scenario, sites c_3 and c_4 mutated twice, yet no recombination events were invoked.

- As Figure 15(b) shows, the evolution of a set of sequences may be explained by multiple mutations at a site, rather than inferring putative recombination events. In this case, the evolutionary history is still a tree, albeit relaxing the infinite-site model to allow recurrent mutations.
- As Figure 16 shows, a gene tree may differ from the species tree due to *lineage sorting* . Informally, lineage sorting happens when two alleles of a gene from two species fail to coalesce, or “merge” at a common ancestral gene, at the divergence time of the two species, and instead they coalesce deeper. We elaborate on this process further below.

Notice that when gene trees disagree with each other, or with the species tree, it is crucial to determine the cause, or causes, of incongruence first, and then use the appropriate reconciliation method. What is needed in practice is a unified, probabilistic framework that, given a set of gene trees, determines the causes of incongruence. It has been argued that a combination of techniques from population genetics and phylogenetics is needed to achieve this goal, particularly to distinguish between reticulate evolutionary events and lineage sorting as probable causes of incongruence [60]. A natural choice for approaching this issue has been to augment the standard *coalescent* theory so as to allow for computing the probabilities of gene trees assuming the presence of events such as horizontal gene transfer.

In a seminal paper, Maddison proposed a framework for inferring the species tree such that both mutations at the nucleotide level and incongruence among gene trees are taken into account [63]. The likelihood of a given species tree, according to [63],

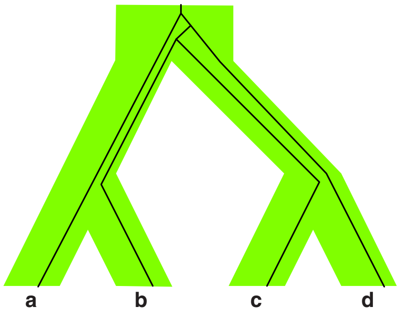


Fig. 16 A gene tree (solid lines) evolving within the branches of the species tree, where the gene tree topology is identical to that of T_2 in Figure 1(b). The gene tree differs from the species tree due to (incomplete) lineage sorting.

is the product, over all loci, of the probability of obtaining the observed sequences at the locus:

$$\prod_{\text{loci}} \sum_{\text{possible gene trees}} [\mathbf{P}(\text{sequences}|\text{gene tree}) \cdot \mathbf{P}(\text{gene tree}|\text{species tree})]. \tag{12}$$

The probability $\mathbf{P}(\text{gene tree}|\text{species tree})$, when deep coalescence is allowed, can be calculated using coalescence theory, as we briefly review in Section 4.1. However, in the most general setting, the species phylogeny may not be a tree. Therefore, an extension to Maddison's framework is necessary to account for reticulate evolutionary events. The ML formulation given in Section 3.3 is similar to Maddison's proposal, but it explicitly models reticulate evolution and ignores lineage sorting. What is needed is an extension to the coalescent to allow for calculating the probability of a gene tree given a species phylogeny assuming any combination of the three discord processes (lineage sorting, reticulate evolution, and gene duplication/loss) could be involved. Preliminary work that simultaneously accounts for lineage sorting and horizontal gene transfer events has been proposed in [100] and another that simultaneously accounts for lineage sorting and hybrid speciation has been proposed in [69]; we review these two in Sections 4.2 and 4.3, respectively.

It is worth mentioning that other approaches for distinguishing reticulate evolution from lineage sorting without explicit modeling of the coalescent process have been introduced. For example, Sang and Zhong proposed a test statistic for distinguishing between lineage sorting and hybridization based on the divergence time of the two parents of a hybrid [89]. However, Holder *et al.* showed later that this statistic fails to reliably distinguish between the two processes [42]. More recently, Holland *et al.* proposed to use *supernetworks* for this task [43].

4.1 Lineage Sorting and the Coalescent

Lineage sorting occurs because of the random contribution of genetic material from each individual in a population to the next generation. Some fail to have offspring while some happen to have multiple offspring. In population genetics, this process was first modeled by R. A. Fisher and S. Wright, in which each gene of the population at a particular generation is chosen independently from the gene pool of the previous generation, regardless of whether the genes are in the same individual or in different individuals. Under the Wright-Fisher model, "the coalescent" considers the process backward in time [44, 57, 97]. That is, the ancestral lineages of genes of interest are traced from offspring to parents. A coalescent event occurs when two (or sometimes more) genes "merge" at the same parent, which is called the most recent common ancestor (MRCA) of the two genes.

The basic process can be treated as follows. Consider a pair of genes at time τ_1 in a randomly mating haploid population. The population size at time τ is denoted by $N(\tau)$. The probability that both genes are from the same parental gene at the previous generation (time τ_1+1) is $1/N(\tau_1+1)$. Therefore, starting at τ_1, the probability that the coalescence between the pair occurs at τ_2 is given by

$$Prob(\tau_2) = \frac{1}{N(\tau_2)} \prod_{\tau=\tau_1+1}^{\tau_2-1} \left(1 - \frac{1}{N(\tau)}\right). \tag{13}$$

When $N(\tau)$ is constant, the probability density distribution (pdf) of the coalescent time (i.e., $t = \tau_2 - \tau_1$) is given by a geometric distribution and can be approximated by an exponential distribution for large N:

$$Prob(t) = \frac{1}{N} e^{-t/N}. \tag{14}$$

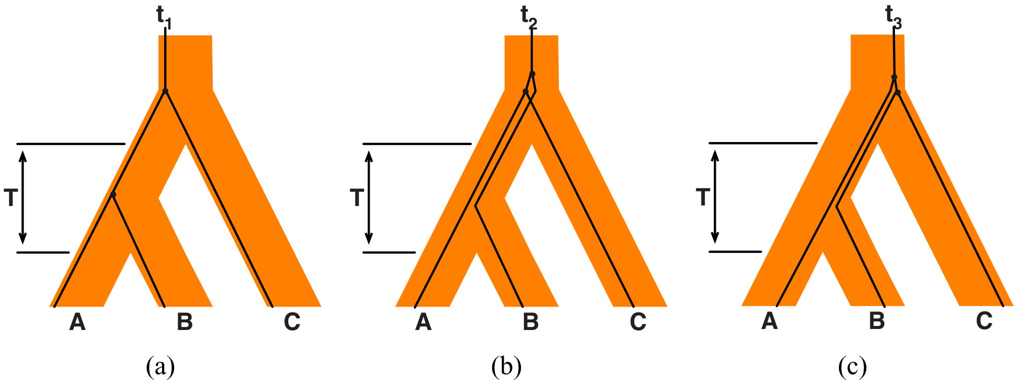


Fig. 17 A species tree on three species A, B, and C. Shown within the branches of the species tree are the three possible gene tree topologies that may result due to different coalescence histories.

Under the three-species model (Figure 17), there are three possible types of gene tree: $(AB)C$, $(AC)B$ and $A(BC)$. Let $Prob$[(AB)C], $Prob$[(AC)B] and $Prob$[A(BC)] be the probabilities of the three types of gene tree. These three probabilities are simply expressed with a continuous time approximation when all populations have equal and constant population sizes, N, where N is large:

$$\mathbf{P}(t_1) = 1 - \frac{2}{3}e^{-T/N}, \tag{15}$$

and

$$\mathbf{P}(t_2) = \mathbf{P}(t_3) = \frac{1}{3}e^{-T/N}. \tag{16}$$

Recently, Rosenberg and colleagues showed that the most likely gene tree may be different from the species tree, when the number of leaves is four or more [19, 88]. It is worth mentioning, however, that when the number of leaves is three, the result does not apply, since the expression in (15) is greater than the expression in (16) for all strictly positive, finite values of T and N.

Observe that in the presence of lineage sorting (in addition to reticulate evolutionary events), the number of gene trees given a (species) phylogenetic network is no longer bounded, as given above by Inequality (1). Rather, the number of possible gene trees now equals the number of possible rooted trees (with the same number of leaves as that of the network). For example, let us consider how the tree $(((a,b),c),d)$ could be one of the gene trees inside the phylogenetic network in Figure 1(a). To obtain this tree, consider the scenario under which b inherits its gene from the a lineage, the genes of c and d fail to coalesce before they reach the root r; instead, c first coalesces with the ancestral gene of a and b, and then the ancestral copy of all three coalesces with that of d. This scenario is illustrated in Figure 18.

4.2 Augmenting the Coalescent with Horizontal Gene Transfer

We now review the model of [100] for extending the coalescent to allow HGT as a cause of incongruence. Suppose that each haploid individual in a population with size N has a lifespan that follows an exponential distribution with mean l. When an individual dies, another individual randomly chosen from the population replaces it to keep the population size constant. In other words, one of the $N-1$ alive lineages is duplicated to replace the dead one. Under the Moran model, the ancestral lineages of individuals of interest can be traced backward in time, and the coalescent time between a pair of individuals follows an exponential distribution with mean $lN/2$ [25, 87]. While phylogeny-based detection of HGT is usually based on quantifying incongruence between a species and a gene tree, the situation becomes more complicated when lineage sorting may be a cause of the incongruence as well.

Consider a model with three species, A, B, and C, in which an HGT event occurs from species B to C, as illustrated in Figure 19. Suppose the MRCA of all three species has a single copy of a gene x. Let a, b and c be the orthologous genes in

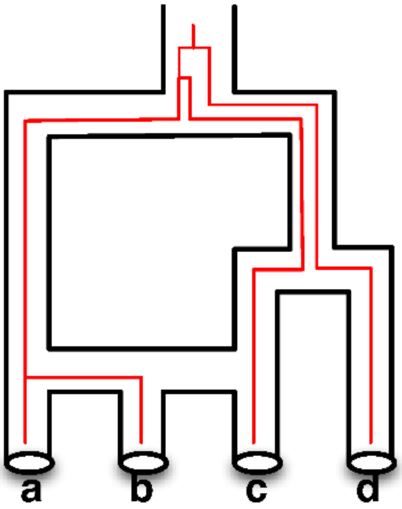


Fig. 18 Illustration of the combined effect of reticulate evolution and lineage sorting. The tubes represent a phylogenetic network in which b is a hybrid taxon (the same as the one in Figure 1(a)), and shown within the tubes is gene tree $(((a,b),c),d)$. Notice that this gene tree cannot be obtained using the **Induce** procedure described in Figure 2, and it is not one of the two trees shown in Figure 1.

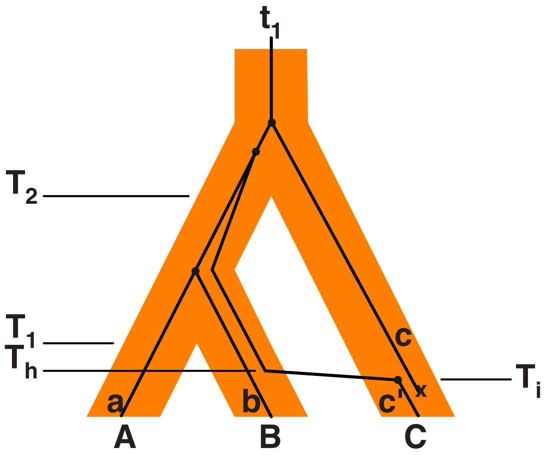


Fig. 19 A three bacterial species model with an HGT event. A demonstration that a congruent tree could be observed even with HGT.

the three species, respectively, whose ancestral gene at the MRCA is x. At time T_h, a gene was transferred from species B and was inserted in a genome in species C at T_i, which is denoted by c'. Since HGT is assumed to be instantaneous at the scale of evolution, in reality, it is always the case that $T_i = T_h$. However, since these times are estimated in practice, it may be the case that $T_h < T_i$. For example, if a gene duplication occurs in lineage b in Figure 19, and one of the two in-paralogs is transferred to c, then the estimated time T_h would be the duplication time, which is earlier than the actual time of the HGT events, T_i.

Following the HGT event, c was physically deleted from the genome, so that each of the three species currently has a single copy of the focal gene. If there is no lineage sorting, the gene tree should be $a(bc')$. Since this tree is incongruent with the species tree, $(AB)C$, we could consider it as an evidence for HGT. However, lineage sorting could also produce the incongruence between the gene tree and species tree without HGT. It is also important to note that lineage sorting, coupled with HGT, could produce a congruent gene tree, as illustrated in Figure 19. Although b and c' have a higher chance to coalesce first, the probability that the first coalescence occurs between a and b or between a and c' may not be negligible especially when $T_1 - T_h$ is short. The probabilities of the three types of gene tree can be formulated under this tri-species model with HGT as illustrated in Figure 19. Here, T_h could exceed T_1; in such a case it can be considered that HGT occurred before the speciation between A and B. Assuming that all populations have equal (constant) population sizes, N, the three probabilities can be obtained modifying (15) and (16):

$$\mathbf{P}[(\mathrm{AB})\mathrm{C}] = \begin{cases} \frac{1}{3}e^{-(T_1-T_h)/N}, & \text{if } T_h \leq T_1 \\ 1-\frac{2}{3}e^{-(T_h-T_1)/N}, & \text{if } T_h > T_1 \end{cases}, \qquad (17)$$

$$\mathbf{P}[(\mathrm{AC})\mathrm{B}] = \begin{cases} \frac{1}{3}e^{-(T_1-T_h)/N}, \text{ if } T_h \leq T_1 \\ \frac{1}{3}e^{-(T_h-T_1)/N}, \text{ if } T_h > T_1 \end{cases}, \qquad (18)$$

and

$$\mathbf{P}[\mathrm{A}(\mathrm{BC})] = \begin{cases} 1-\frac{2}{3}e^{-(T_1-T_h)/N}, & \text{if } T_h \leq T_1 \\ \frac{1}{3}e^{-(T_h-T_1)/N}, & \text{if } T_h > T_1 \end{cases}. \qquad (19)$$

4.3 Augmenting the Coalescent with Hybrid Speciation

We now review the model of [69] for extending the coalescent to allow hybrid speciation as a cause of incongruence, using the scenario depicted in Figure 20 as an example. The issue at hand is, given a collection of genes whose trees may be incongruent, whether their incongruence due to hybrid speciation or lineage sorting. In the former case, their reconciliation would result in the phylogenetic network depicted by the wide bands in Figure 20. However, as the time T between the MRCA of any two of the species and the MRCA of all three becomes smaller, the probability of gene tree disagreement due to lineage sorting increases.

Let a, b, and c be three orthologous genes randomly sampled from the three species A, B, and C, respectively, where B is a hybrid of A and C. The model of Meng and Kubatko assumes that when a gene b is arbitrarily selected from species B, then its most recent common ancestor occurs with species A with probability γ and with species C with probability $1-\gamma$. These two possible trees are t_1 and t_2, respectively, discussed in the caption of Figure 20. Once one of these two trees is selected, the model treats the tree as a species tree and allows the coalescent process

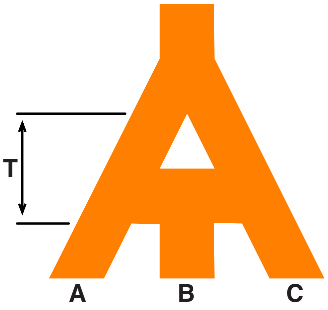


Fig. 20 A phylogenetic network representing a hybrid speciation event involving species A and C, and producing species B. The two possible (alternative) species trees are $t_1 = ((A,B),C)$ and $t_2 = (A,(B,C))$.

to operate for that gene.[7] Using Equations (15) and (16) above for calculating the probabilities of gene trees given a species tree, and assuming t_1 as a species tree, we have

$$\mathbf{P}[((\mathrm{A},\mathrm{B}),\mathrm{C})] = 1 - \tfrac{2}{3}e^{-T/N},$$
$$\mathbf{P}[((\mathrm{A},\mathrm{C}),\mathrm{B})] = \mathbf{P}[(\mathrm{A},(\mathrm{B},\mathrm{C}))] = \tfrac{1}{3}e^{-T/N}.$$

Assuming t_2 as the species tree, we have

$$\mathbf{P}[(\mathrm{A},(\mathrm{B},\mathrm{C}))] = 1 - \tfrac{2}{3}e^{-T/N},$$
$$\mathbf{P}[((\mathrm{A},\mathrm{C}),\mathrm{B})] = \mathbf{P}[((\mathrm{A},\mathrm{B}),\mathrm{C})] = \tfrac{1}{3}e^{-T/N}.$$

The question is to determine, given a collection of genes sampled from the genomes of the three species, whether the evolutionary history of the three species is the phylogenetic network in Figure 20, the species tree t_1, or the species tree t_2. One way of answering this question is to estimate the probability γ. If $\gamma = 1$, then the evolutionary history of the three species is the species tree t_1. If $\gamma = 0$, then the evolutionary history of the three species is the species tree t_2. If $0 < \gamma < 1$, then the evolutionary history is the phylogenetic network shown in Figure 20, with the proportions of the genome of B inherited from A and C are γ and $(1-\gamma)$, respectively.

Meng and Kubatko provided a maximum likelihood estimation of the parameters γ and T, as well as a Bayesian estimation technique [69]. We briefly review the main points of the maximum likelihood estimation. Let $\mathscr{G} = \{gt_1, gt_2, \ldots, gt_k\}$ be an *i.i.d.* sample of gene trees, where gt_i is the tree of gene i, sampled so that their topologies are independent and follow the hybridization model described in Figure 20. The likelihood function for a given phylogenetic network with a specified location for the hybrid speciation event (as shown in Figure 20) is given by:

$$L(\gamma, T|\mathscr{G}) = \prod_{i=1}^{k} \mathbf{P}(gt_i|\gamma, T) = \prod_{i=1}^{k} [\gamma \mathbf{P}(gt_i|t_1, T) + (1-\gamma)\mathbf{P}(gt_i|t_2, T)] \qquad (20)$$

[7] Notice the similarity between this and the probability of a tree as given by Equation (9).

Notice that this formula is a special case of Formula (10) (when taken for multiple genes) given in Definition 0.5. Formula (10) is defined for networks with any number of hybrid speciation events, and the parameter M_T in the formula is a generalization of the pairs (t_1, T) and (t_2, T) in Formula (20), since M_T is the model, which includes the tree topology, its branch lengths, and the model of evolution.

The question now becomes one of estimating the parameters γ and T that maximize the likelihood function and determining, based on these (particularly γ), whether the phylogenetic network or tree is the evolutionary history of the species, and, if the latter, which of the two (t_1 or t_2).

5 Exercises

Here we give a set of exercises for the reader to gain a better understanding of evolutionary phylogenetic networks and issues related to their reconstruction and evaluation.

1. Show an example of a set $\mathscr{T}$ of trees, with $|\mathscr{T}| = 2$, and a minimal network N that reconciles both trees in $\mathscr{T}$ such that $\mathscr{T} \neq \mathscr{T}(N)$.
2. Figure 11 shows one minimal ARG for the given sequence data set. Draw all other minimal ARGs.
3. a. Show a phylogenetic network N with $|\mathscr{T}(N)| = 2^k$, where k is the number of network-nodes in N.
   b. Show a phylogenetic network N with $|\mathscr{T}(N)| < 2^k$, where k is the number of network-nodes in N.
4. Show two trees, each with nine leaves, whose SPR distance is 3, and for which the number of minimal phylogenetic networks that reconcile the two trees is 27. (Hint: Consider trees with three clades, each clade with three leaves, and each clade requires a single SPR move.)
5. For each of the two characters labeling the leaves of the network N in Figure 12(b), show a tree in $\mathscr{T}(N)$ on which the character is compatible, by also showing the labeling of internal nodes of the tree.
6. Using the illustration in Figure 18, describe one coalescence scenario for each of the possible gene trees that are induced by the phylogenetic network in Figure 1 assuming lineage sorting could occur.

6 Further Reading

An excellent resource on phylogenetic networks is *Who is Who in Phylogenetic Networks* [31], which, as of the date of writing this manuscript, catalogs 264 publications and 34 software tools dedicated to phylogenetic networks. There have been several recent detailed surveys of phylogenetic reconstruction methods [32, 47, 48, 61, 64, 72], some of which identify their similarities and differences.

Recently, several results have appeared on measures for comparing phylogenetic network topologies and quantifying their dissimilarities; we refer the reader to [3, 8, 9, 10, 11, 14, 15, 70, 76, 79]. Further, some proposals have been made on representing phylogenetic networks for I/O operations using an *extended Newick*, or eNewick, format; e.g., see [12, 13, 71, 101].

Acknowledgements The author would like to thank the editors for the invitation to contribute this chapter, Joe Felsenstein, Lenny Heath, and Mike Steel for extensive and insightful comments that helped improve the manuscript significantly, and Guohua Jin, Laura S. Kubatko, Kevin Liu, Richard Manning, Francesc Rosselló, and Yun Yu for helpful comments on the text. The illustration in Figure 18 and its accompanying discussion are the result of a discussion with Laura S. Kubatko and James Degnan. This work was supported in part by DOE grant DE-FG02-06ER25734, NSF grant CCF-0622037, and grant R01LM009494 from the National Library of Medicine. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the DOE, NSF, National Library of Medicine or the National Institutes of Health.

References

1. Allen, B., Steel, M.: Subtree transfer operations and their induced metrics on evolutionary trees. Annals of Combinatorics **5**, 1–13 (2001)
2. Baroni, M., Grunewald, S., Moulton, V., Semple, C.: Bounding the number of hybridisation events for a consistent evolutionary history. J. Math. Biol. **51**, 171–182 (2005)
3. Baroni, M., Semple, C., Steel, M.: A framework for representing reticulate evolution. Annals of Combinatorics **8**(4), 391–408 (2004)
4. Beiko, R., Hamilton, N.: Phylogenetic identification of lateral genetic transfer events. BMC Evolutionary Biology **6** (2006)
5. Bergthorsson, U., Adams, K., Thomason, B., Palmer, J.: Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature **424**, 197–201 (2003)
6. Bergthorsson, U., Richardson, A., Young, G., Goertzen, L., Palmer, J.: Massive horizontal transfer of mitochondrial genes from diverse land plant donors to basal angiosperm Amborella. Proc. Nat'l Acad. Sci., USA **101**, 17,747–17,752 (2004)
7. Bordewich, M., Semple, C.: On the computational complexity of the rooted subtree prune and regraft distance. Annals of Combinatorics **8**, 409–423 (2004)
8. Cardona, G., Llabrés, M., Rosselló, F., Valiente, G.: A distance metric for a class of tree-sibling phylogenetic networks. Bioinformatics **24**(13), 1481–1488 (2008)
9. Cardona, G., Llabrés, M., Rosselló, F., Valiente, G.: Metrics for phylogenetic networks I: Generalizations of the robinson-foulds metric. IEEE/ACM Transactions on Computational Biology and Bioinformatics **6**(1), 1–16 (2009)
10. Cardona, G., Llabrés, M., Rosselló, F., Valiente, G.: Metrics for phylogenetic networks II: Nodal and triplets metrics. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2009)
11. Cardona, G., Llabrés, M., Rosselló, F., Valiente, G.: On Nakhleh's latest metric for phylogenetic networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2009). To appear
12. Cardona, G., Rosselló, F., Valiente, G.: Extended Newick: It is time for a standard representation of phylogenetic networks. BMC Bioinformatics **9**, 532 (2008)
13. Cardona, G., Rossello, F., Valiente, G.: A Perl package and an alignment tool for phylogenetic networks. BMC Bioinformatics **9**(1), 175 (2008)
14. Cardona, G., Rosselló, F., Valiente, G.: Tripartitions do not always discriminate phylogenetic networks. Mathematical Biosciences **211**(2), 356–370 (2008)

15. Cardona, G., Rosselló, F., Valiente, G.: Comparison of tree-child phylogenetic networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2009). To appear
16. Chor, B., Tuller, T.: Maximum likelihood of evolutionary trees is hard. Proc. 9th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB05) pp. 296–310 (2005)
17. Darwin, C.: On the origin of species by means of natural selection. J. Murray, London (1859)
18. Day, W.: Computationally difficult parsimony problems in phylogenetic systematics. Journal of Theoretical Biology **103**, 429–438 (1983)
19. Degnan, J., Rosenberg, N.: Discordance of species trees with their most likely gene trees. PLoS Genetics **2**, 762–768 (2006)
20. Doolittle, W.: Lateral genomics. Trends in Biochemical Sciences **24**(12), M5–M8 (1999)
21. Doolittle, W.: Phylogenetic classification and the universal tree. Science **284**, 2124–2129 (1999)
22. Edwards, S.: Is a new and general theory of molecular systematics emerging? Evolution **63**(1), 1–19 (2009)
23. Ellstrand, N., Whitkus, R., Rieseberg, L.: Distribution of spontaneous plant hybrids. Proc. Nat'l Acad. Sci., USA **93**(10), 5090–5093 (1996)
24. Estabrook, G., McMorris, F.: When are two qualitative taxonomic characters compatible? J. Math. Biosci. **4**, 195–200 (1977)
25. Ewens, W.: Mathematical Population Genetics. Springer-Verlag, Berlin (1979)
26. Felsenstein, J.: Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology **27**, 401–410 (1978)
27. Felsenstein, J.: Alternative methods of phylogenetic inference and their interrelationship. Systematic Zoology **28**, 49–62 (1979)
28. Felsenstein, J.: Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. **17**, 368–376 (1981)
29. Fitch, W.: Toward defining the course of evolution: Minimum change for a specified tree topology. Syst. Zool. **20**, 406–416 (1971)
30. Foulds, L., Graham, R.: The Steiner problem in phylogeny is NP-complete. Adv. Appl. Math. **3**, 43–49 (1982)
31. Gambette, P.: Who is who in phylogenetic networks: Articles, authors and programs. http://www.lirmm.fr/∼gambette/PhylogeneticNetworks/
32. Gemeinholzer, B.: Phylogenetic networks. In: B.H. Junker, F. Schreiber (eds.) Analysis of Biological Networks, pp. 255–282. John Wiley and Sons Ltd (2008)
33. Goloboff, P.: Calculating SPR distances between trees. Cladistics **24**, 591–597 (2007)
34. Griffiths, R., Marjoram, P.: An ancestral recombination graph. In: P. Donnelly, S. Tavare (eds.) Progress in Population Genetics and Human Evolution, *IMA Volumes in Mathematics and its Applications*, vol. 87, pp. 257–270. Springer-Verlag, Berlin (1997)
35. Gusfield, D.: Efficient algorithms for inferring evolutionary trees. Networks **21**, 19–28 (1991)
36. Gusfield, D., Bansal, V., Bafna, V., Song, Y.: A decomposition theory for phylogenetic networks and incompatible characters. Journal of Computational Biology **14**, 1247–1272 (2007)
37. Gusfield, D., Eddhu, S., Langley, C.: Efficient reconstruction of phylogenetic networks with constrained recombination. In: Proceedings of Computational Systems Bioinformatics (CSB 03) (2003)
38. Hallett, M., Lagergren, J.: Efficient algorithms for lateral gene transfer problems. In: Proc. 5th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB01), pp. 149–156. ACM Press, New York (2001)
39. Hao, W., Golding, G.: Patterns of bacterial gene movement. Mol. Biol. Evol. **21**(7), 1294–1307 (2004)
40. Hein, J.: Reconstructing evolution of sequences subject to recombination using parsimony. Math. Biosciences **98**, 185–200 (1990)
41. Hein, J.: A heuristic method to reconstruct the history of sequences subject to recombination. J. Mol. Evol. **36**, 396–405 (1993)
42. Holder, M., Anderson, J., Holloway, A.: Difficulties in detecting hybridization. Systematic Biology **50**(6), 978982 (2001)

43. Holland, B., Benthin, S., Lockhart, P., Moulton, V., Huber, K.: Using supernetworks to distinguish hybridization from lineage-sorting. BMC Evolutionary Biology **8**, 202 (2008)
44. Hudson, R.: Properties of the neutral allele model with intergenic recombination. Theor. Popul. Biol. **23**, 183–201 (1983)
45. Hudson, R., Kaplan, N.: Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111**, 147–164 (1985)
46. Humphries, P., Semple, C.: Note on the hybridization number and subtree distance in phylogenetics. Applied Mathematics Letters (2009). In press
47. Huson, D.H.: Split networks and reticulate networks. In: O. Gascuel, M. Steel (eds.) Reconstructing Evolution, New Mathematical and Computational Advances, pp. 247–276. Oxford University Press (2007)
48. Huson, D.H., Bryant, D.: Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution **23**(2), 254–267 (2006)
49. Jin, G., Nakhleh, L., Snir, S., Tuller, T.: Efficient parsimony-based methods for phylogenetic network reconstruction. Bioinformatics **23**, e123–e128 (2006). Proceedings of the European Conference on Computational Biology (ECCB 06)
50. Jin, G., Nakhleh, L., Snir, S., Tuller, T.: Maximum likelihood of phylogenetic networks. Bioinformatics **22**(21), 2604–2611 (2006)
51. Jin, G., Nakhleh, L., Snir, S., Tuller, T.: Inferring phylogenetic networks by the maximum parsimony criterion: A case study. Molecular Biology and Evolution **24**(1), 324–337 (2007)
52. Jin, G., Nakhleh, L., Snir, S., Tuller, T.: A new linear-time heuristic algorithm for computing the parsimony score of phylogenetic networks: Theoretical bounds and empirical performance. In: I. Mandoiu, A. Zelikovsky (eds.) Proceedings of the International Symposium on Bioinformatics Research and Applications, *Lecture Notes in Bioinformatics*, vol. 4463, pp. 61–72 (2007)
53. Kanj, I., Nakhleh, L., Than, C., Xia, G.: Seeing the trees and their branches in the network is hard. Theoretical Computer Science **401**, 153–164 (2008)
54. Kanj, I., Nakhleh, L., Xia, G.: The compatibility of binary characters on phylogenetic networks: Complexity and parameterized algorithms. Algorithmica **51**, 99–128 (2008)
55. Kimura, M.: The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61**, 893–903 (1969)
56. Kimura, M., Crow, J.: The number of alleles that can be maintained in a finite population. Genetics **49**, 725–738 (1964)
57. Kingman, J.F.C.: The coalescent. Stochast. Proc. Appl. **13**, 235–248 (1982)
58. Kurland, C., Canback, B., Berg, O.: Horizontal gene transfer: A critical view. Proc. Nat'l Acad. Sci., USA **100**(17), 9658–9662 (2003)
59. Lathrop, G.: Evolutionary trees and admixture: Phylogenetic inference when some populations are hybridized. Ann. Hum. Genet. **46**, 245–255 (1982)
60. Linder, C., Rieseberg, L.: Reconstructing patterns of reticulate evolution in plants. American Journal of Botany **91**, 1700–1708 (2004)
61. Linder, C.R., Moret, B.M.E., Nakhleh, L., Warnow, T.: Network (reticulate) evolution: Biology, models, and algorithms. In: The Pacific Symposium on Biocomputing (2004)
62. MacLeod, D., Charlebois, R., Doolittle, F., Bapteste, E.: Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. BMC Evolutionary Biology **5** (2005)
63. Maddison, W.: Gene trees in species trees. Systematic Biology **46**(3), 523–536 (1997)
64. Makarenkov, V., Kevorkov, D., Legendre, P.: Phylogenetic network construction approaches. In: Applied Mycology and Biotechnology, pp. 61–97 (2006)
65. Mallet, J.: Hybridization as an invasion of the genome. TREE **20**(5), 229–237 (2005)
66. Mallet, J.: Hybrid speciation. Nature **446**, 279–283 (2007)
67. McClilland, M., Sanderson, K., Clifton, S., Latreille, P., Porwollik, S., Sabo, A., Meyer, R., Bieri, T., Ozersky, P., McLellan, M., Harkins, C., Wang, C., Nguyen, C., Berghoff, A., Elliott, G., Kohlberg, S., Strong, C., Du, F., Carter, J., Kremizki, C., Layman, D., Leonard, S., Sun, H., Fulton, L., Nash, W., Miner, T., Minx, P., Delehaunty, K., Fronick, C., Magrini, V., Nhan,

M., Warren, W., Florea, L., Spieth, J., Wilson, R.: Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *salmonella enterica* that cause typhoid. Nature Genetics **36**(12), 1268–1274 (2004)

68. Meacham, C.: Theoretical and computational considerations of the compatibility of qualitative taxonomic characters. NATO ASI Series **G1 on Numerical Taxonomy** (1983)
69. Meng, C., Kubatko, L.: Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. Theoretical Population Biology **75**(1), 35–45 (2009)
70. Moret, B., Nakhleh, L., Warnow, T., Linder, C., Tholse, A., Padolina, A., Sun, J., Timme, R.: Phylogenetic networks: Modeling, reconstructibility, and accuracy. IEEE/ACM Transactions on Computational Biology and Bioinformatics **1**(1), 13–23 (2004)
71. Morin, M., Moret, B.: NetGen: Generating phylogenetic networks with diploid hybrids. Bioinformatics **22**(15), 1921–1923 (2006)
72. Morrison, D.A.: Networks in phylogenetic analysis: new tools for population biology. International Journal of Parasitology **35**, 567–582 (2005)
73. Mower, J., Stefanovic, S., Young, G., Palmer, J.: Gene transfer from parasitic to host plants. Nature **432**, 165–166 (2004)
74. Nakamura, Y., Itoh, T., Matsuda, H., Gojobori, T.: Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nature Genetics **36**(7), 760–766 (2004)
75. Nakhleh, L.: Phylogenetic networks. Ph.D. thesis, The University of Texas at Austin (2004)
76. Nakhleh, L.: A metric on the space of reduced phylogenetic networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2009). To appear
77. Nakhleh, L., Ringe, D., Warnow, T.: Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. LANGUAGE, Journal of the Linguistic Society of America **81**(2), 382–420 (2005)
78. Nakhleh, L., Ruths, D., Wang, L.: RIATA-HGT: A fast and accurate heuristic for reconstructing horizontal gene transfer. In: L. Wang (ed.) Proceedings of the Eleventh International Computing and Combinatorics Conference (COCOON 05), pp. 84–93 (2005). LNCS #3595
79. Nakhleh, L., Sun, J., Warnow, T., Linder, R., Moret, B., Tholse, A.: Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. In: Proceedings of the 8th Pacific Symposium on Biocomputing, pp. 315–326. World Scientific Pub. (2003)
80. Noor, M., Feder, J.: Speciation genetics: Evolving approaches. Nature Review Genetics **7**, 851–861 (2006)
81. Ochman, H., Lawrence, J., Groisman, E.: Lateral gene transfer and the nature of bacterial innovation. Nature **405**(6784), 299–304 (2000)
82. Posada, D., Crandall, K.: The effect of recombination on the accuracy of phylogeny estimation. J. Mol. Evol. **54**(3), 396–402 (2002)
83. Posada, D., Crandall, K., Holmes, E.: Recombination in evolutionary genomics. Annu. Rev. Genet. **36**, 75–97 (2002)
84. Pupko, T., Pe'er, I., Shamir, R., Graur, D.: A fast algorithm for joint reconstruction of ancestral amino-acid sequences. Mol. Biol. Evol. **17**(6), 890–896 (2000)
85. Rieseberg, L., Baird, S., Gardner, K.: Hybridization, introgression, and linkage evolution. Plant Molecular Biology **42**(1), 205–224 (2000)
86. Rieseberg, L., Carney, S.: Plant hybridization. New Phytologist **140**(4), 599–624 (1998)
87. Rosenberg, N.: Gene genealogies. In: C. Fox, J.B. Wolf (eds.) Evolutionary Genetics: Concepts and Case Studies, chap. 15. Oxford Univ. Press University Press (2005)
88. Rosenberg, N., Tao, R.: Discordance of species trees with their most likely gene trees: The case of five taxa. Systematic Biology **57**, 131–140 (2008)
89. Sang, T., Zhong, Y.: Testing hybridization hypotheses based on incongruent gene trees. Systematic Biology **49**(3), 422434 (2000)
90. Song, Y.: Properties of subtree-prune-and-regraft operations on totally-ordered phylogenetic trees. Annals of Combinatorics **10**, 129–146 (2006)

91. Song, Y., Ding, Z., Gusfield, D., Langley, C., Wu, Y.: Algorithms to distinguish the role of gene-conversion from single-crossover recombination in the derivation of SNP sequences in populations. Journal of Computational Biology **14**, 1273–1286 (2007)
92. Song, Y., Hein, J.: Parsimonious reconstruction of sequence evolution and haplotype blocks: Finding the minimum number of recombination events. In: Proc. 3rd Int'l Workshop Algorithms in Bioinformatics (WABI03), vol. 2812, pp. 287–302. Springer-Verlag (2003)
93. Song, Y., Hein, J.: On the minimum number of recombination events in the evolutionary history of DNA sequences. Journal of Mathematical Biology **48**, 160–186 (2004)
94. Song, Y., Hein, J.: Constructing minimal ancestral recombination graphs. Journal of Computational Biology **12**, 147–169 (2005)
95. Steel, M., Penny, D.: Parsimony, likelihood, and the roles of models in molecular phylogenetics. Mol. Biol. Evol. **17**, 839–850 (2000)
96. Strimmer, K., Moulton, V.: Likelihood analysis of phylogenetic networks using directed graphical models. Mol. Biol. Evol. **17**, 875–881 (2000)
97. Tajima, F.: Evolutionary relationship of DNA sequences in finite populations. Genetics **105**, 437–460 (1983)
98. Than, C., Jin, G., Nakhleh, L.: Integrating sequence and topology for efficient and accurate detection of horizontal gene transfer. In: Proceedings of the Sixth RECOMB Comparative Genomics Satellite Workshop, *Lecture Notes in Bioinformatics*, vol. 5267, pp. 113–127 (2008)
99. Than, C., Nakhleh, L.: SPR-based tree reconciliation: Non-binary trees and multiple solutions. In: Proceedings of the Sixth Asia Pacific Bioinformatics Conference (APBC), pp. 251–260 (2008)
100. Than, C., Ruths, D., Innan, H., Nakhleh, L.: Confounding factors in HGT detection: Statistical error, coalescent effects, and multiple solutions. Journal of Computational Biology **14**(4), 517–535 (2007)
101. Than, C., Ruths, D., Nakhleh, L.: PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics **9**, 322 (2008)
102. Wang, L., Zhang, K., Zhang, L.: Perfect phylogenetic networks with recombination. Journal of Computational Biology **8**(1), 69–78 (2001)
103. Welch, R., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E., Liou, S., Boutin, A., Hackett, J., Stroud, D., Mayhew, G., Rose, D., Zhou, S., Schwartz, D., Perna, N., Mobley, H., Donnenberg, M., Blattner, F.: Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *escherichia coli*. Proc. Nat'l Acad. Sci., USA **99**(26), 17,020–17,024 (2002)
104. Willson, S.: Reconstruction of certain phylogenetic networks from the genomes at their leaves. Journal of Theoretical Biology **252**, 338–349 (2008)
105. Wu, Y.: A practical method for exact computation of subtree prune and regraft distance. Bioinformatics **25**(2), 190–196 (2009)