Syst. Biol. 60(2):138–149, 2011

DOI:10.1093/sysbio/syq084
Advance Access publication on January 19, 2011

Coalescent Histories on Phylogenetic Networks and Detection of Hybridization Despite Incomplete Lineage Sorting

YUN YU[1], CUONG THAN[2], JAMES H. DEGNAN[3], LUAY NAKHLEH[1,*]

[1]*Department of Computer Science, Rice University, 6100 Main Street, Houston, TX 77005, USA;*
[2]*Department of Human Genetics, University of Michigan, 1241 East Catherine Street, Ann Arbor, MI 48109, USA; and*
[3]*Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand;*
**Correspondence to be sent to: Department of Computer Science, Rice University, 6100 Main Street, Houston, TX 77005, USA; E-mail: nakhleh@cs.rice.edu.*

Received 2 November 2009; reviews returned 9 February 2010; accepted 30 August 2010
Associate Editor: Laura Kubatko

Abstract.—Analyses of the increasingly available genomic data continue to reveal the extent of hybridization and its role in the evolutionary diversification of various groups of species. We show, through extensive coalescent-based simulations of multilocus data sets on phylogenetic networks, how divergence times before and after hybridization events can result in incomplete lineage sorting with gene tree incongruence signatures identical to those exhibited by hybridization. Evolutionary analysis of such data under the assumption of a species *tree* model can miss all hybridization events, whereas analysis under the assumption of a species *network* model would grossly overestimate hybridization events. These issues necessitate a paradigm shift in evolutionary analysis under these scenarios, from a model that assumes a priori a single source of gene tree incongruence to one that *integrates* multiple sources in a unifying framework. We propose a framework of coalescence within the branches of a *phylogenetic network* and show how this framework can be used to detect hybridization despite incomplete lineage sorting. We apply the model to simulated data and show that the signature of hybridization can be revealed as long as the interval between the divergence times of the species involved in hybridization is not too small. We reanalyze a data set of 106 loci from 7 in-group Saccharomyces species for which a species tree with no hybridization has been reported in the literature. Our analysis supports the hypothesis that hybridization occurred during the evolution of this group, explaining a large amount of the incongruence in the data. Our findings show that an integrative approach to gene tree incongruence and its reconciliation is needed. Our framework will help in systematically analyzing genomic data for the occurrence of hybridization and elucidating its evolutionary role. [Coalescent history; incomplete lineage sorting; hybridization; phylogenetic network.]

Hybridization is believed to play an important role in the speciation and evolutionary innovations of several groups of plant and animal species (Arnold 1997; Mallet 2007). Whether hybridization is polypoloid or diploid, the evolutionary histories of different marker alleles in a hybrid species can take different paths through the two parental populations. This evolutionary fact is the basis for a large class of phylogeny-based methods for detecting hybridization (or, reticulate evolution in general) in a group of taxa. These methods compare the evolutionary histories of different genomic regions and take incongruence in their individual evolutionary histories to indicate hybridization (e.g., see Nakhleh 2010 for a recent survey of these methods).

A major factor that confounds the performance, in terms of the accuracy of the inferred species evolutionary history, of hybridization detection methods is that gene tree incongruence may be caused by other factors, such as incomplete lineage sorting (also referred to as deep coalescence) (Maddison 1997). Indeed, several recent studies have reported on massive amounts of incongruence in various data sets due to incomplete lineage sorting (Syring et al. 2005; Pollard et al. 2006; Kuo et al. 2008; Than, Sugino, et al. 2008; Cranston et al. 2009). Consequently, recognizing the need to reconcile the incongruence in multilocus data sets is giving rise to a new paradigm for inferring species phylogenies (Edwards 2009). Indeed, several methods have been developed for inferring species trees from multilocus data despite incomplete lineage sorting. In the *concatenation approach*, the sequences from multiple loci are concatenated, and the resulting "supergene" data set is analyzed using traditional phylogenetic methods, such as maximum parsimony or maximum likelihood (Rokas et al. 2003). Methods for summarizing gene trees have also been used in this context. One way to reconcile the gene trees is by taking their majority consensus (Kuo et al. 2008). Another is the "democratic vote" method, which entails taking the tree topology occurring with the highest frequency among all gene trees as the species tree. Finally, a class of methods that explicitly model the coalescent process has been introduced. This includes Bayesian inference methods (Edwards et al. 2007; Liu and Pearl 2007; Heled and Drummond 2010), a maximum likelihood method (Kubatko et al. 2009), and parsimony methods (Maddison 1997; Maddison and Knowles 2006; Than, Sugino, et al. 2008; Than and Nakhleh 2009). Several surveys of species tree inference methods and issues related to inferring species trees despite incomplete lineage sorting are available (Rannala and Yang 2008; Degnan and Rosenberg 2009; Liu et al. 2009)

Nonetheless, just as methods for detecting hybridization events typically assume that hybridization is the sole cause of gene tree incongruence and infer a species *network*, these latter methods assume that incomplete lineage sorting is the sole cause of incongruence in the data and infer a species *tree*. However, as hybridization may occur between closely related species, incongruence among evolutionary histories of genomic regions

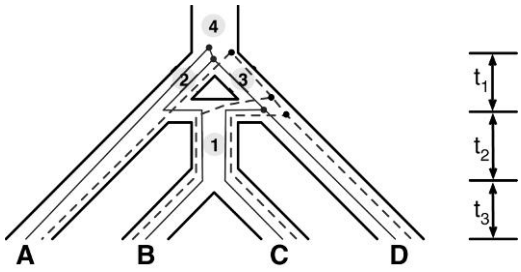

FIGURE 1. Phylogenetic networks and deep coalescence. Two valid coalescent histories that give rise to gene tree (A,(B,(C,D))) are shown within the branches of a phylogenetic network of four species. The effective population sizes, times t_1 and t_2, as well as the probability γ of a gene being inherited through hybridization, all determine the distribution of gene genealogies in the four species.

may be partly due to incomplete lineage sorting, and distinguishing between the two factors is hard under these conditions (Mallet 2005).

Figure 1 shows a scenario of an evolutionary history of four taxa that involves hybridization, as well as incomplete lineage sorting. If the divergence time between the two species prior to hybridization (time t_1) and the time between hybridization and a subsequent speciation (time t_2) are long enough, we expect that the two gene tree topologies ((A,(B,C)),D) and (A,((B,C),D)) would appear with the highest frequencies. If time t_2 is long enough, yet t_1 is not, we expect a third gene tree topology, namely ((A,D),(B,C)), to appear with a high frequency. If time t_1 is long enough, yet t_2 is not, then we expect eight gene tree topologies to appear with higher frequencies than other topologies: ((A,(B,C)),D), (A,((B,C),D)), ((A,B),(C,D)), ((A,C),(B,D)), (((A,B),C),D), (((A,C),B),D), (A,(B,(C,D))), and (A,(C,(B,D))). When neither t_1 nor t_2 is long enough, all 15 (rooted) binary tree topologies on four leaves may be observed as the shapes of the gene genealogies. Another factor that determines the distribution of the observed gene genealogies is γ, the probability of alleles in the ancestral population of B and C being inherited from A (as opposed to being inherited from D).

Existing methods for phylogenetic network reconstruction (e.g., see Nakhleh 2010 for a survey) ignore incomplete lineage sorting as a cause of incongruence, which amounts to the assumption that t_1 and t_2 are too long for deep coalescence to occur. On the other hand, methods for inferring species trees in the presence of lineage sorting ignore hybridization, which amounts to the assumption that $\gamma = 0$ or $\gamma = 1$ for the network in Figure 1. When either assumption holds, existing methods have been shown to compute accurate estimates of the species phylogenies. However, under most circumstances, such assumptions cannot be made a priori. Otherwise, applying hybridization detection methods may result in overestimating the amount of hybridization, and inferring species trees results in missing any hybridization that may have occurred. Therefore, a more appropriate model is a phylogenetic network that allows for deep coalescence events because such a structure allows for simultaneously capturing vertical and horizontal inheritance of genetic material (Linder and Rieseberg 2004).

We are aware of four works that attempt to address this issue of simultaneous modeling of incomplete lineage sorting and horizontal evolution. Than et al. (2007) introduced a stochastic framework for detecting horizontal gene transfer, given a species tree and a gene tree, in the presence of lineage sorting. Later, Meng and Kubatko (2009) and Kubatko (2009) introduced another coalescent-based framework for detecting hybridization in the presence of lineage sorting. These methods were tested on simulated data, and their performance was analyzed. Holland et al. (2008) proposed using supernetworks (a generalization of the concept of supertrees) to distinguish hybridization from incomplete lineage sorting. More recently, Joly et al. (2009) introduced a statistical framework for the same task, which distinguishes hybridization from incomplete lineage sorting based on the genetic distances between sequences. This framework entails conducting coalescent-based simulations for testing the null hypothesis of only incomplete lineage sorting and no hybridization.

In this paper, we study the effect of the parameters outlined in Figure 1 on the detectability of hybridization in the presence of incomplete lineage sorting and develop a parsimony-based method for detecting hybridization in the presence of lineage sorting. The method performs very well, except for the cases where times t_1 and t_2 are very small, when incomplete lineage sorting is too rampant for a hybridization signal to be detected. We reanalyze the 106-locus yeast data set of Rokas et al. (2003) using our method and detect hybridization. This data set has been well studied and analyzed using different methods (Rokas et al. 2003; Edwards et al. 2007; Than and Nakhleh 2009). All these studies and analyses have converged on a single species tree. Nonetheless, our analysis produced a hypothesis on the evolutionary history of this well-studied data set that supports a hybridization event involving *Saccharomyces kudriavzevii*, *S. bayanus*, and the clade containing *S. mikatae*, *S. cerevisiae*, and *S. paradoxus*. These results are in agreement with those of Bloomquist and Suchard (2010), as we discuss below.

METHODS

In this section, we describe the hybridization model we consider, as well as a parsimony-based inference method. In the next section, we discuss how the model and inference method can be extended to more general cases.

The Model

In this paper, we consider the 4-taxon, single-allele hybridization model depicted in Figure 1, which includes a hybridization event involving A, D, and the clade (B,C) and allows for deep coalescence events to

take place within the branches of the species network. This scenario is more complex than that investigated in previous studies of hybridization that allow for incomplete lineage sorting (Than et al. 2007; Holland et al. 2008; Meng and Kubatko 2009; Joly et al. 2009) in that it allows for divergence after hybridization. This, in turn, enables us to study the effect of the divergence time between the two "parents" (t_1 in Figure 1) as well as the time between the hybridization and subsequent divergence (t_2 in Fig. 1) on the ability to detect hybridization.

We denote by γ the probability that an allele in the ancestral population of B and C is inherited from A (and $1-\gamma$ is the probability that an allele is inherited from D). Under this model, the gene tree topology G can be viewed as a random variable whose probability mass function is $P_{N,\lambda}(G=g)$, where λ is a vector of the branch lengths of the phylogenetic network N. We list all 15 4-taxon binary gene tree topology probabilities for the network N in Table 1, where the probability of a gene tree is the sum of the terms in the first column with coefficients in the appropriate column for the gene tree. For example, the probability of the gene tree (((A,D),B),C) is

$$c \cdot \frac{1}{9} g_{22}(t_2)(g_{22}(t_1))^2 + d \cdot \frac{1}{18} g_{22}(t_2) g_{33}(t_1) \qquad (1)$$

where $c = \gamma(1-\gamma)$ and $d = \gamma^2 + (1-\gamma)^2$, and $g_{ij}(t)$ is the probability that i lineages coalesce into j lineages within time t. Expressions for $g_{ij}(t)$ are given by Tavaré (1984) and Nordborg (1998). The two terms in Equation (1) correspond to the cases that either the B and C lineages follow different paths or the same path up through the network "before" coalescing with the (A,D) lineage above the root of the network.

Parsimony-based Inference: Minimizing Deep Coalescences on Networks

The above formulation of the model naturally gives rise to a maximum likelihood framework for estimating γ and the branch lengths of phylogenetic network N. The likelihood function is

$$L(\gamma, \lambda | N, \mathcal{G}, \Theta) = \prod_{i=1}^{\ell} P_{N,\lambda}(G = g_i),$$

where $\mathcal{G}$ is the collection of ℓ input gene trees and Θ is a vector of population sizes in N. Although this model is similar to that of Meng and Kubatko (2009) and Kubatko (2009), it differs in that it departs from the assumption of *parental species trees*, in which a gene tree evolves in a species tree displayed by the network, because this concept does not generalize in a natural way to cases where multiple alleles are sampled or where a hybridization is followed by population divergence.

Nonetheless, in this paper, we study the power of parsimonious reconciliations of gene trees at detecting hybridization in the presence of lineage sorting. In particular, we focus on the question of estimating γ by extending the *minimize deep coalescences* (MDC) criterion of Maddison (1997) to phylogenetic networks.

Given a *valid coalescence history* (Degnan and Salter 2005) of a gene tree g within the branches of species tree T, we can count the number of extra lineages resulting from this coalescence coalescent history. The minimum number thus obtained from any valid coalescence history of g within the branches of T is considered the number of extra lineages, and denoted by $XL(T,g)$. We can generalize this notion to a set $\mathcal{G}$ of gene trees as

$$XL(T, \mathcal{G}) = \sum_{g \in \mathcal{G}} XL(T, g). \qquad (2)$$

Using this definition, the problem of inferring the species tree under the MDC criterion, which we call the MDC-T problem, is defined as follows.

Definition 1 (The MDC-T Problem)

Input: Set $\mathcal{G}$ of gene trees.

TABLE 1. The probability distribution of 15 gene trees given the species network

	Trees								
Term	(((A,B),C),D) (((A,C),B),D)	(((A,B),D),C) (((A,C),D),B)	(((A,D),B),C) (((A,D),C),B)	(((B,D),A),C) (((C,D),A),B)	(((B,D),C),A) (((C,D),B),A)	((A,B),(C,D)) ((A,C),(B,D))	(((B,C),A),D)	(((B,C),D),A)	((A,D),(B,C))
$g_{22}(t_2)(g_{21}(t_1))^2$	0	0	0	0	0	c	0	0	0
$\frac{1}{3}g_{22}(t_2)g_{21}(t_1)g_{22}(t_1)$	c	c	0	c	c	$2c$	0	0	0
$\frac{1}{9}g_{22}(t_2)(g_{22}(t_1))^2$	c	c	c	c	c	$2c$	c	c	$2c$
$\frac{1}{3}g_{22}(t_2)g_{31}(t_1)$	a^2	0	0	0	b^2	0	a^2	b^2	0
$\frac{1}{9}g_{22}(t_2)g_{32}(t_1)$	a^2	a^2	0	b^2	b^2	d	d	d	d
$\frac{1}{18}g_{22}(t_2)g_{33}(t_1)$	d	d	d	d	d	$2d$	d	d	$2d$
$g_{21}(t_2)g_{21}(t_1)$	0	0	0	0	0	0	a	b	0
$\frac{1}{3}g_{21}(t_2)g_{22}(t_1)$	0	0	0	0	0	0	1	1	1

Note: We take $a=\gamma$, $b=(1-\gamma)$, $c=\gamma(1-\gamma)$, and $d=\gamma^2+(1-\gamma)^2$. Each column gives the coefficients of the terms in the first column in terms of the hybridization parameter γ.

Output: $ST = \text{argmin}_T XL(T, \mathcal{G})$.

Maddison and Knowles (2006) developed a heuristic for solving the MDC-T problem. Than and Nakhleh (2009) recently proposed the first exact solution to the problem and demonstrated its efficiency, showing that it runs in seconds on data sets with tens of taxa and thousands of loci. In both studies, MDC was shown to perform very well for both biological and synthetic data sets.

As described above, the MDC criterion, as proposed by Maddison (1997), allows for inferring a species *tree*, assuming only deep coalescence events as the cause of gene tree incongruence. The concept of a *valid coalescent history* can be extended to the domain of phylogenetic networks in a straightforward manner. When tracing the coalescent within the branches of a tree, each allele follows a unique path from a leaf in the tree to its root. However, in a network, more than one path may be followed by some alleles from a leaf to the root. For example, an allele from species A in Figure 1 follows the unique path from the leaf labeled by A to the root. On the other hand, an allele from species B can follow one of two paths: it either goes "left" at the hybridization node, or it goes "right." Once the notion of valid coalescent histories is modified as such, the concept of extra lineages and the MDC criterion apply, unchanged, to phylogenetic networks. We now show how to estimate the value of γ under the MDC criterion on networks.

Although there are 15 binary gene trees on four taxa, we have to consider, as well, nonbinary gene trees for inference, because reconstructed gene trees need not be fully resolved. For easy description of valid coalescent histories within the branches of the network, we adapt the Newick format for trees to include the branch number in the network of Figure 1 on which a clade of the gene tree coalesces. The 26 possible gene tree topologies $g_1, \ldots, g_{26}$, along with valid coalescent histories that define the gene trees under the MDC criterion are shown in Figure 2. For example, under the MDC criterion, the valid coalescent history that gives rise to the gene tree (((A,C),D),B) is the one where the A and C alleles coalesce on branch 2 of the network, then (A,C) and D alleles coalesce on branch 4, and the remaining two alleles coalesce on branch 4; hence, we write (((A,C)2,D)4,B)4.

Under the MDC criterion, we define the two *inheritance* vectors $\mathbf{l}$ and $\mathbf{r}$, which denote the number of alleles in the ancestral hybrid population (looking backward in time, this is the ancestral population of B and C that does not yet involve a hybridization event) that were *necessarily* inherited from A (which equals the number of coalescence events on branch 2 in the network) and D (which equals the number of coalescence events on branch 3 of the network), respectively. The length of each of the vectors is 26, with one entry for each of the 26 gene trees topologies. For the phylogenetic network in Figure 1, we have:

- $\mathbf{l} = [2,1,2,1,0,0,1,0,0,0,0,0,0,1,1,0,2,0,0,0,1,1,0,0,0,0,0]$.
- $\mathbf{r} = [0,0,0,0,0,0,0,0,1,1,2,1,2,1,1,0,0,0,0,0,2,0,0,0,0,1,1,0]$.

A value of 0 in vector $\mathbf{l}$ ($\mathbf{r}$) denotes that, under the MDC criterion, no coalescence events occurred on the branch to the left (right) of the root for that topology; such an event does not contribute to the γ estimate. A value of 1 in $\mathbf{l}$ ($\mathbf{r}$) corresponds to the case where exactly one allele from the hybrid population coalesces first with A (D). A value of 2 in $\mathbf{l}$ ($\mathbf{r}$) corresponds to the case where the ancestral hybrid population has two alleles (i.e., the alleles from B and C have not coalesced within time t_2), and both of them coalesce first with A (D). Let us illustrate this concept with the gene tree g_{12} = (A,(B,(C,D))). Figure 1 shows two valid coalescent histories for this gene tree. Under the MDC criterion, the one shown in solid lines is not the one that defines the number of extra lineages. The "optimal" valid coalescent history in this case is the one shown in dashed lines, which has two coalescent events on the branch to the right of the root (first, C and D coalesce, and then their most recent common ancestor [MRCA] coalesce with B). This coalescent history h provides two pieces of information that are used to estimate γ:

1. h implies that the two alleles from B and C did not coalesce in the ancestral hybrid population.
2. Further, h implies that *both* alleles "went right"; that is, both alleles were inherited from D.

Based on these two facts, gene tree g_{12} contributes 0 to the vector $\mathbf{l}$, and 2 to the vector $\mathbf{r}$. Indeed, $\mathbf{l}(12) = 0$ and $\mathbf{r}(12) = 2$.

Using these two vectors, as well as the frequencies of each of the 26 topologies in the input, we can get an estimate of γ. Let $\mathbf{f}$ be a vector of the frequencies of the ℓ input gene trees, that is, $f(i)$ is the frequency of gene tree g_i in the input, and $\sum_{i=1}^{26} f(i) = \ell$. Then, the estimate $\hat{\gamma}$ is computed as the proportion of lineages that go left assuming that each gene tree is obtained by the coalescent history that has the smallest number of extra lineages:

$$\hat{\gamma} = \begin{cases} 0 & \text{if } (\mathbf{l}\cdot\mathbf{f} + \mathbf{r}\cdot\mathbf{f} = 0) \\ \frac{\mathbf{l}\cdot\mathbf{f}}{\mathbf{l}\cdot\mathbf{f}+\mathbf{r}\cdot\mathbf{f}} & \text{otherwise} \end{cases} \tag{3}$$

where $\mathbf{a}\cdot\mathbf{b}$ denotes the dot product of two vectors $\mathbf{a}$ and $\mathbf{b}$.

The Case of an Unknown Network

Notice that thus far we have assumed a given network N. However, in practice, the input to the problem is a set $\mathcal{G}$ of trees, and the goal is to estimate the phylogenetic network (possibly a tree), along with the value $\hat{\gamma}$ for each hybrid node. The space of network topologies is huge and grows very fast with the number of species under consideration. Although an exhaustive search of all networks is possible for very small numbers of taxa (e.g., up to 5), this task becomes prohibitive for larger numbers of species.

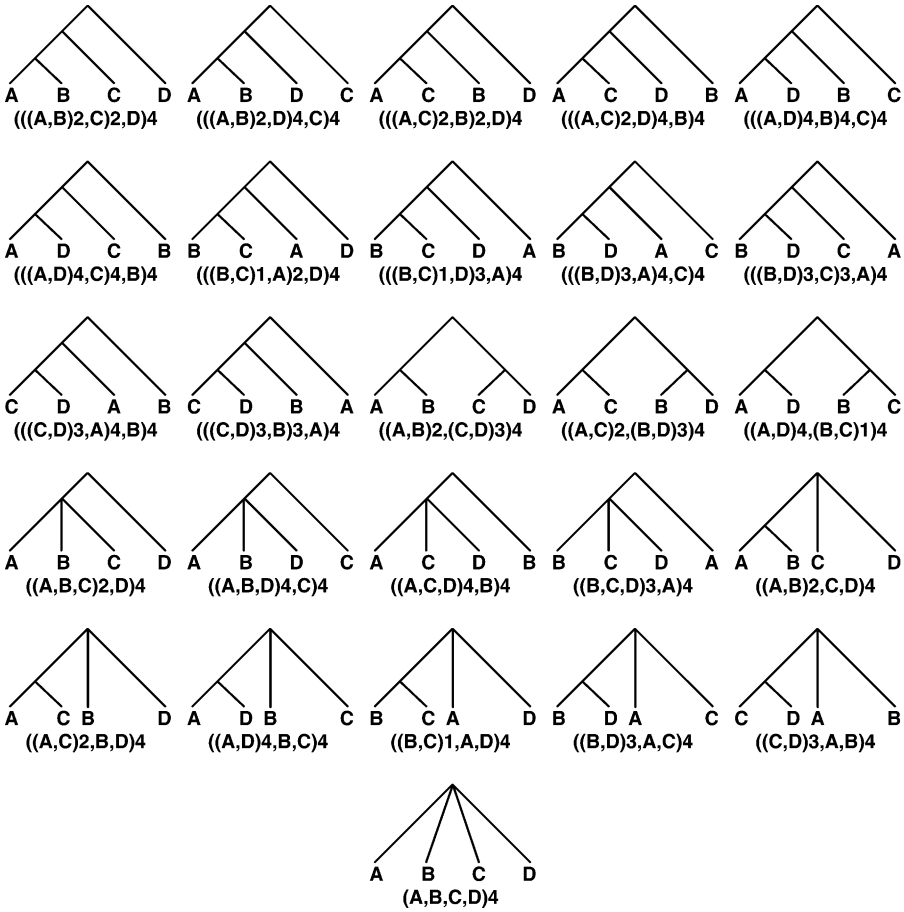

FIGURE 2. The 26 possible gene tree topologies with four leaves. The numbers within the Newick format description of each tree correspond to the branch numbers in the phylogenetic network of Figure 1.

Another significant problem that arises when searching for a "good" network concerns the problem of overfitting; without controlling the number of reticulations (hybrid nodes) in the network, identifying an optimal network becomes a trivial task because we can always find a network N that reconciles the entire set of gene trees with no extra lineages (e.g., Nakhleh 2010). This overfitting phenomenon is an issue that plagues phylogenetic network reconstruction in general; the more reticulations in the network, the better it fits the data. Therefore, without a close inspection of the improvement to fitting the data, one may end up with a network that grossly overestimates the amount of hybridization (Nakhleh 2010). Even worse, one can always find a network that reconciles each gene tree without resulting in any deep coalescence events. This is analogous to the problem of inferring phylogenetic networks to model sequence evolution, where one can always find a network under which each site in the sequences evolves with no homoplasy (Jin et al. 2007). Kubatko (2009) has recently addressed this issue in the case of maximum likelihood inference of hybridization.

To address the computational and overfitting issues, we propose the following method for detecting hybridization in the presence of lineage sorting from a set $\mathcal{G}$ of gene trees.

Procedure Trees2Network

1. Solve MDC-T on $\mathcal{G}$ to compute an optimal species tree T^*.
2. Compute the set
$$\mathcal{T} = \left\{ T : \frac{XL(T,\mathcal{G}) - XL(T^*,\mathcal{G})}{XL(T^*,\mathcal{G})} \leq \theta\% \right\} \quad (4)$$
of trees that are within θ% of optimality.
3. Infer a minimal phylogenetic network N such that $\mathcal{T} \subseteq \mathcal{T}(N)$, where $\mathcal{T}(N)$ is the set of all trees inside N.

In other words, this procedure guides the network space search by the set $\mathcal{T}$ of trees, thus reducing the computational complexity. Furthermore, it ameliorates the overfitting problem by focusing on a set $\mathcal{T}$ of candidate trees, rather than continually augmenting the network. Although this procedure is still computationally intensive, it is expected to be much faster in practice than an exhaustive search. We show below how this

heuristic performs on the yeast data set for different values of θ.

Data

To study the performance of our method, we conducted coalescent-based simulations under the probabilistic model described above. This model allows us to simulate gene trees within the branches of the phylogenetic network of Figure 1, where the parameters t_1, t_2, and γ control the frequencies of all 15 gene tree topologies. We have implemented the simulation procedure in PhyloNet (Than, Ruths, et al. 2008), which can be run with the command line:

sim_GT in Network t_1 t_2 γ ℓ

where, in addition to t_1, t_2, and γ, the user specifies ℓ, which is the number of gene trees to be simulated. Alternatively, a Matlab implementation of the simulation procedure can be downloaded from PhyloNet's website (http://bioinfo.cs.rice.edu/phylonet).

In our simulations, we varied the times t_1 and t_2 to take on the values 0.5, 1, 2, and 4 in terms of coalescent units (time in generations normalized by population size), ranging from the very short (and hence extensive deep coalescence) to the very long (and hence almost no deep coalescence), respectively. Furthermore, we used values 0, 0.1, 0.2, 0.3, 0.4, and 0.5 for γ, to simulate cases with the amount of hybridization ranging from none to equal contribution of both parents, respectively. We varied the number of sampled loci, ℓ, so that we could study the impact of the size of the input data on our method's ability to detect hybridization. In our study, we considered $\ell \in \{25, 50, 100, 250, 500, 1000\}$. In all of our simulations, we used only a single allele per species. For every combination of values of t_1, t_2, γ, and ℓ, we simulated 30 sets of gene trees. Furthermore, to study the effect of error in the estimated gene trees, we used the Seq-gen program of Rambaut and Grassly (1997) to simulate the evolution of DNA sequences of length 1000 under the model of Jukes and Cantor (1969) down each of the gene trees. We reconstructed gene trees from these sequence alignments under the maximum likelihood criterion in the program PAUP* assuming the Jukes–Cantor model.

Furthermore, we reanalyzed the yeast data set of Rokas et al. (2003). The yeast data set of Rokas et al. (2003) contains seven Saccharomyces species, *S. cerevisiae* (Scer), *S. paradoxus* (Spar), *S. mikatae* (Smik), *S. kudriavzevii* (Skud), *S. bayanus* (Sbay), *S. castellii* (Scas), *S. kluyveri* (Sklu), and the outgroup fungus *Candida albicans* (Calb). Rokas et al. (2003) identified 106 genes, which are distributed throughout the *S. cerevisiae* genome on all 16 chromosomes and comprise about 2% of the predicted genes. For each gene, they reconstructed its tree using the maximum likelihood and maximum parsimony methods. Among the 106 trees, more than 20 different gene tree topologies were observed. Rokas et al. (2003) inferred the species tree using the concatenation method on the sequences of the 106 genes. The resulting tree had 100% bootstrap support for each of its branches; this tree topology is shown in Figure 3a. Furthermore, various studies of the same data set, using different criteria and methods, have inferred this same tree as the species tree best supported by the 106 gene trees (Edwards et al. 2007; Than and Nakhleh 2009).

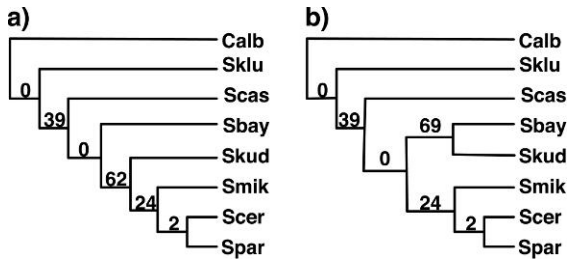

FIGURE 3. The species tree of the yeast data set, as proposed by Rokas et al. a) The single optimal tree under the MDC criterion for the data set. The number of extra lineages resulting from reconciling all 106 gene trees within the branches of this tree is 127. b) The best suboptimal tree under the MDC criterion. The number of extra lineages resulting from reconciling all 106 gene trees within the branches of this tree is 134, which is just seven extra lineages away from the optimal value of 127 achieved by the tree in (a). The number on a branch indicates the number of extra lineages along that branch once all 106 gene trees are reconciled within the branches of the tree.

RESULTS AND DISCUSSION

Results on Simulated Data

For the simulated data, we mainly investigated the effect of the times t_1 and t_2 and the number of gene trees sampled on the estimation of γ from the data, using Equation (3). Our hypothesis was that the longer both t_1 and t_2 were, and the larger the number of sampled gene trees, the more accurate the γ estimate would be. Further, we hypothesized that as the value of γ increased, a better signal of hybridization would emerge, and hence, be detected. Figure 4 shows the results of the γ estimates on the true gene trees, whereas Figure 5 shows the results of the γ estimates on the reconstructed gene trees. Although we show only a subset of the results, the trends hold for the remaining data.

As Figures 4 and 5 show, the method does very well in the case of $t_1 = t_2 = 4$. In this case, we expect that alleles from B and C coalesce first within t_2, then their ancestral allele goes toward A or D, depending on the value of γ, and finally coalesce with A or D within time t_1. In other words, in this case, we expect the majority of the input gene trees to have one of the two topologies ((A,(B,C)),D) and (A,((B,C),D)), and that the proportion of the two tree topologies provides a good indication of the value of γ. Indeed, both figures show that the value of $\hat{\gamma}$ is almost identical to γ in the case of $t_1 = t_2 = 4$, whether the true or reconstructed gene trees are used, and even for the case when only 25 gene trees are used. An inspection of the actual probabilities of the gene tree topologies, given in Figure 6, shows that in the case of

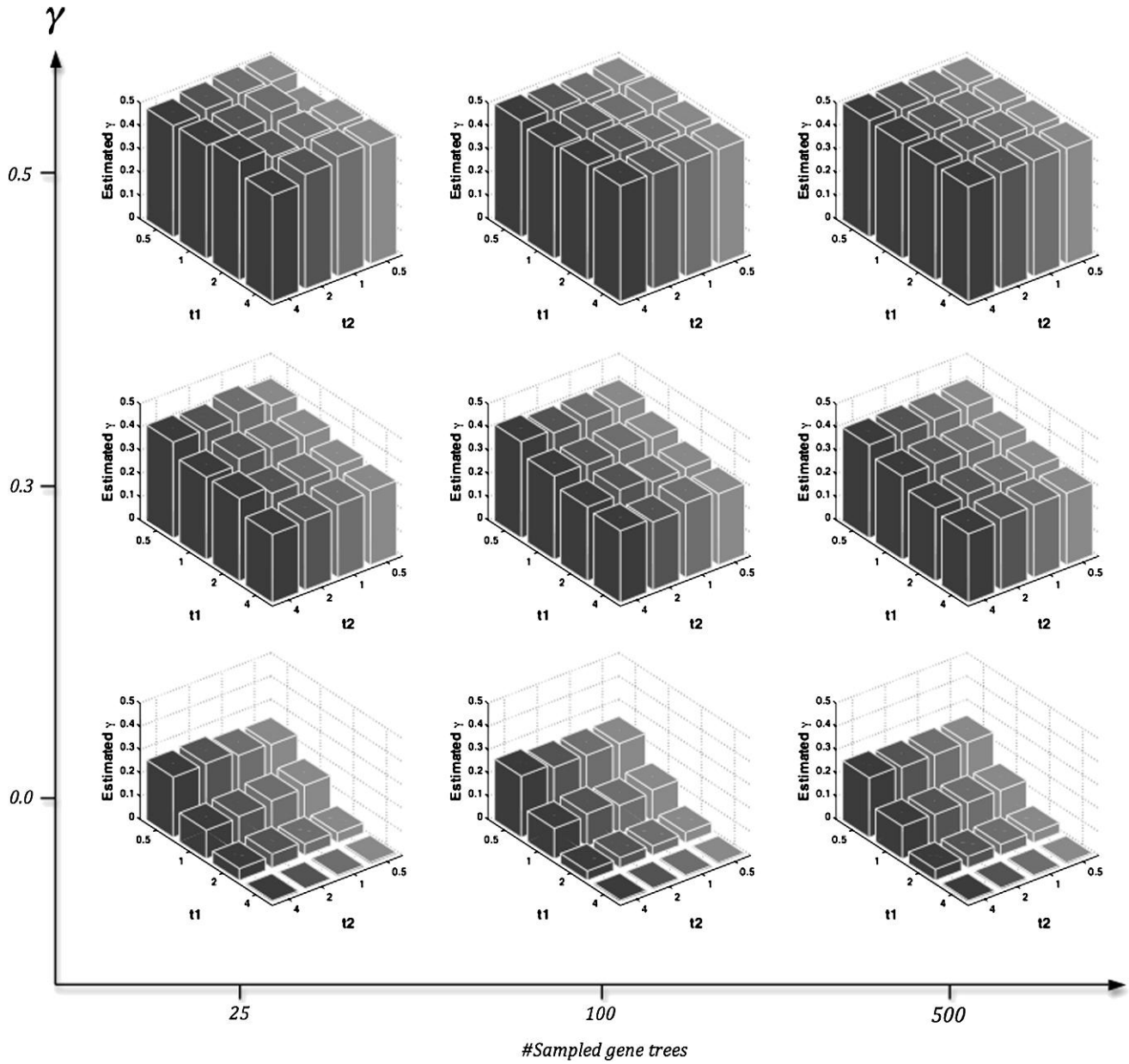

FIGURE 4. Estimated γ ($\hat{\gamma}$ in Equation (3)) from true gene trees. The variance of the estimates is lower than 0.001 in most cases, and reaches a maximum value of 0.008.

$t_1 = t_2 = 4$, the only two tree topologies that have probabilities much higher than 0 are g_7 and g_8, which are ((A,(B,C)),D) and (A,((B,C),D)). Furthermore, the relative probabilities of the two trees reflects the value of γ.

In the case when $t_1 = 0.5$ and $t_2 = 4$, Figure 6 shows that a third tree topology, in addition to g_7 and g_8, is likely to appear with a probability higher than 0; this is $g_{15} = ((A,D),(B,C))$, which reflects that no coalescence events involving A, D, and (B,C) occurred below the root. In this case, and based on the definition of the vectors **l** and **r**, tree g_{15} does not contribute to the estimate of γ, and hence we obtain an overestimation of γ, as shown in Figures 4 and 5.

Figure 6 shows that for the case of $t_1 = t_2 = 0.5$, almost all 15 gene tree topologies occur with nonnegligible probability. In this case, we do not expect a parsimony-based criterion to obtain a good estimate of γ, which is what we observe in Figures 4 and 5.

An important observation is that the value of t_1 has much more of an effect on the γ estimate than t_2. In fact, for a fixed value of t_1, the value of $\hat{\gamma}$ seems to remain unchanged with changing value of t_2. And for all different values of γ we examined, we have $\hat{\gamma} \approx \gamma$ as long as t_1 is long enough. This actually makes sense under the MDC criterion where alleles are always assumed to coalesce as close to the species' MRCA as possible. When

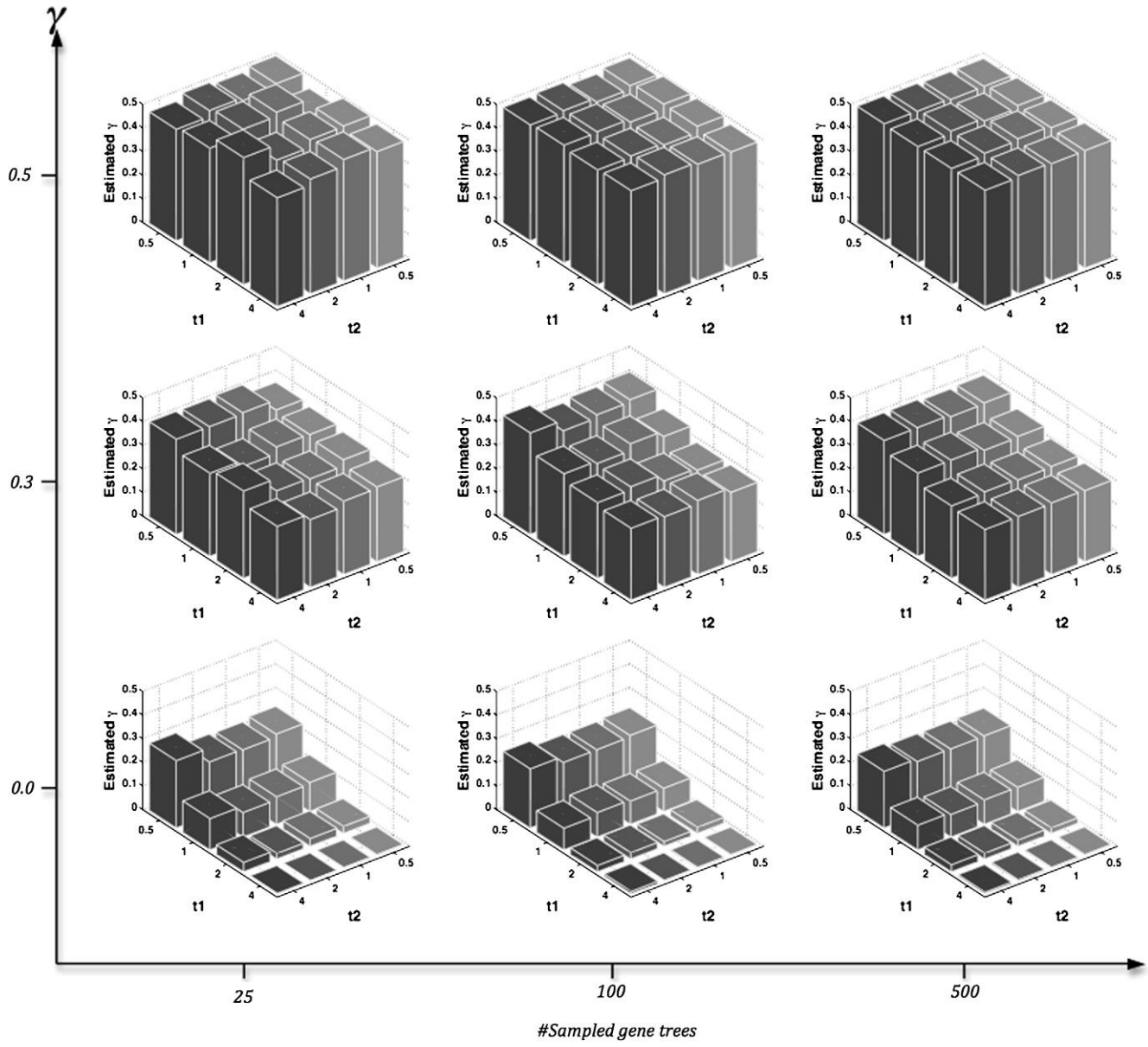

FIGURE 5. Estimated γ ($\hat{\gamma}$ in Equation (3)) from reconstructed gene trees. The variance of the estimates is lower than 0.001 in most cases, and reaches a maximum value of 0.008.

$t_1 = 4$, almost all gene trees have B and C coalesced with A or D below the root. So the topologies of gene trees do reflect the value of γ to some extent. Of course, as t_2 decreases, the amount of deep coalescence may increase, but its effect on the γ estimate is not big. For example, when $t_2 = 0.5$ and $t_1 = 4$, a gene tree ((A,(B,C)),D) may be the outcome of a valid coalescent history in which the B and C alleles do not coalesce within time t_2; instead, they both go left, coalesce within time t_1, and then their ancestral allele coalesces with A. In this scenario, the "true" coalescent history contributes 2 to the inheritance vector **l** and 0 to **r**. However, under the MDC criterion, the valid coalescent history that has the B and C alleles coalescing within time t_2 is taken as the one contributing the extra lineages. As a result, this gene tree contributes 1 to vector **l** and 0 to vector **r**. In other words, we get a slight underestimation of γ in this case, but in general, t_2 has little effect on γ estimate.

On the other hand, when t_1 decreases, the amount of deep coalescence increases significantly, and many possible gene tree topologies appear, thus decreasing the power to estimate γ. Let us consider the case of $t_1 = 0.5$, $t_2 = 4$ and $\gamma = 0$. Since t_2 is long enough and t_1 is too short, it is highly likely that the B and C alleles coalesce within time t_2 and their ancestral allele goes right but does not coalesce with D's allele within time t_1. Finally, when the three alleles from A, D, and the ancestor of (B,C) enter the root population, any pair of

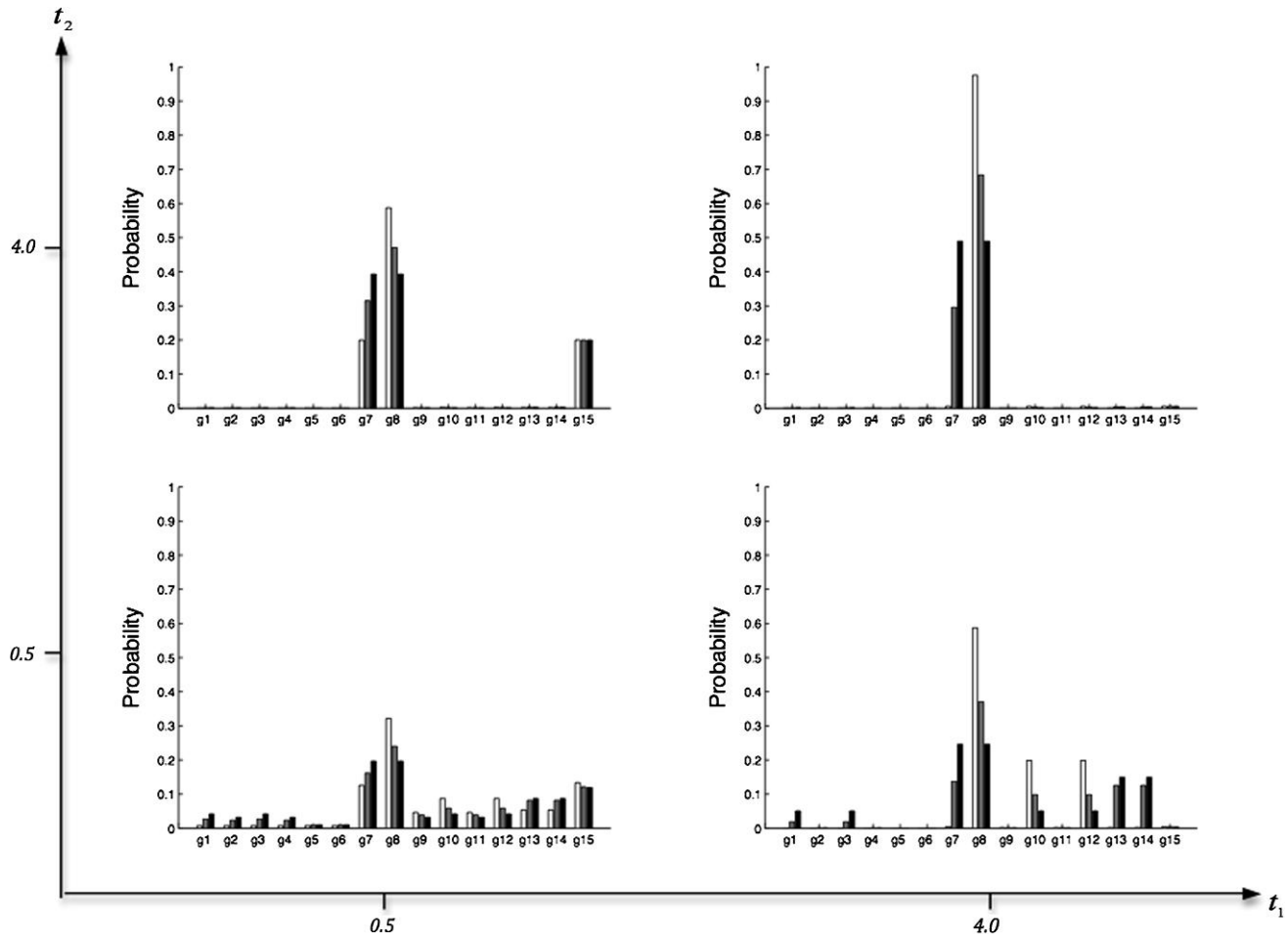

FIGURE 6. Theoretical distribution of 15 gene tree topologies. White bars: $\gamma = 0$; gray bars: $\gamma = 0.3$; black bars: $\gamma = 0.5$.

coalescence events is possible, including one that gives rise to the gene tree topology ((A,(B,C)),D). In this case, the gene tree contributes 1 to the inheritance vector **l**, even though it should contribute 0.

To summarize this issue, the fact that the true coalescent history is not necessarily the one that defines the (optimal) number of extra lineages results in wrong estimates of γ. This issue becomes worse as t_1 decreases, whereas it does not seem to be much affected by the value of t_2.

Two other observations from Figures 4 and 5 are in order. First, for the parameters tested in simulation, the γ estimates become more accurate as more loci are sampled, with good performance even for the smallest sample size (25 loci). Second, there is not much difference in the performance of the method between true and reconstructed gene trees, indicating robustness of the method to errors in the gene tree estimates.

Still, the performance of the method suffers when both times t_1 and t_2 decrease significantly because the amount of deep coalescence increases significantly. What the results indicate is that when the extent of deep coalescences becomes massive, a network becomes a much better representation of the data, even in the absence of any hybridization. In this case, we would expect a more sophisticated approach, such as a stochastic method that also attempts to estimate times, population sizes, etc., would do much better than a parsimony-based method such as the one we present here. It may be possible to improve the performance of the parsimony-based method by coupling it with coalescent-based simulations under the null hypothesis of no hybridization. However, once again, the performance of such an approach would heavily depend on the accuracy of population parameter estimates.

Results on the Yeast Data Set

For our analysis of the yeast data set, we reconstructed the gene trees using a maximum parsimony heuristic, and used our method (Than and Nakhleh 2009) to infer the optimal species tree under the MDC criterion. There was a single optimal tree, which is identical to that proposed by Rokas et al. (2003), and is

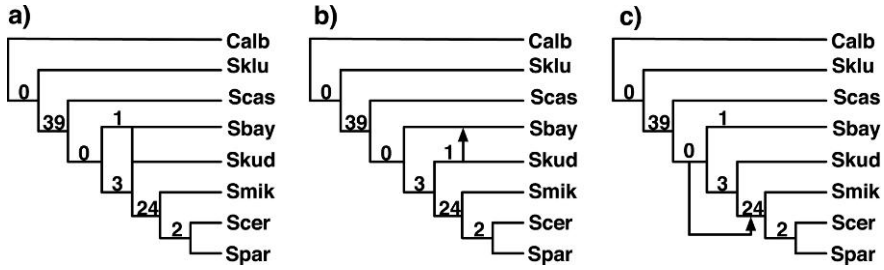

FIGURE 7. Three hybridization scenarios for the yeast data set. Each of the networks requires 69 extra lineages to reconcile all 106 gene trees, and depicts a slightly different hybridization scenario. The number on a branch indicates the number of extra lineages along that branch once all 106 gene trees are reconciled within the branches of the network.

shown in Figure 3a. This tree, ST, results in 127 extra lineages when optimally reconciling all 106 gene trees in the data set. Furthermore, the number of extra lineages incurred on each of the next six suboptimal species tree candidates ST_1, ST_2, ST_3, ST_4, ST_5, and ST_6 were 134, 163, 170, 186, 191, and 193. Using the Trees2Network heuristic above, with $\theta = 5$, results in the single optimal tree given in Figure 3a. In other words, using $\theta = 5$ indicates no hybridization. However, if we take $\theta = 6$ (or, for any $5 < \theta < 28$), we obtain the set $\mathcal{T}$ of the two trees shown in Figure 3, and the Trees2Network heuristic computes the three networks shown in Figure 7, each of which is a minimal network for the two trees in $\mathcal{T}$. For $\theta \geq 28$, the set $\mathcal{T}$ includes additional trees, and the network contains more than a single hybridization event. However, in this case, the improvement in the number of extra lineages is very small.

The network in Figure 7a illustrates a scenario in which hybridization occurred between *S. bayanus* and the clade of *S. mikatae*, *S. cerevisiae*, and *S. paradoxus* to give rise to hybrid species *S. kudriavzevii*. The network in Figure 7b illustrates a scenario in which hybridization occurred between *S. kudriavzevii* and *S. bayanus*. The network in Figure 7c seemingly illustrates a scenario in which hybridization occurred between an ancestor of all five species and the clade of *S. mikatae*, *S. cerevisiae*, and *S. paradoxus*. Such a scenario may at first sound implausible given that it violates the natural constraint that hybridization involves two species that *coexist in time*. However, this is not necessarily the case, as this type of violation can be explained through incomplete taxon sampling or extinction (Nakhleh 2010). This scenario can be explained, for example, if the hybridization had occurred between the clade of *S. mikatae*, *S. cerevisiae*, and *S. paradoxus* and a sibling of the clade of all five species that was not sampled, or became extinct.

A striking point about all three networks is that they all induced exactly the same pair of trees. Ignoring the hybridization event in all of them, we obtain the optimal tree shown in Figure 3a. However, in each of the three networks, if lineages take the alternative route through the network suggested by the hybridization event, we obtain the tree shown in Figure 3b. Interestingly, this tree is second only to the optimal tree, shown in Figure 3a, as shown by the MDC analysis of Than and Nakhleh (2009) and the Bayesian analysis of Edwards et al. (2007). Furthermore, it is very close, in terms of the optimality value, to the optimal one: only seven extra lineages separate the two.

Notice that the differences between the two trees in Figures 3a and 3b are in the relationships between the three groups: (I) *S. kudriavzevii*, (II) *S. bayanus*, and (III) the clade of *S. mikatae*, *S. cerevisiae*, and *S. paradoxus*. Whereas the optimal *tree* groups II with III, each of the three hybridization scenarios shown in Figure 7 indicates that hybridization could be a better supported hypothesis than that given by the optimal tree. Of the 106 gene trees, 65 have the clade ((*S. paradoxus*, *S. cerevisiae*), *S. mikatae*), and 38 have the clade (*S. bayanus*, *S. kudriavzevii*). Furthermore, each of the 106 loci in all five species have coalesced at the MRCA of these five species, as indicated by the value 0 on the branch above the MRCA in all three scenarios in Figure 7.

Finally, it is worth mentioning that these results are identical to those found using the stochastic framework of Bloomquist and Suchard (2010). The only difference is that Bloomquist and Suchard excluded *Sklu* and *Calb* from their analysis; however, these two species are not involved in the hypothesized hybridization scenario. Several studies have reported on the presence of hybridization in yeast (e.g., Gonzalez et al. 2007, 2008). In particular, Dunn and Sherlock (2008) have recently reported on a hybridization between *S. cerevisiae* and *S. bayanus*-related yeasts to form *S. pastorianus*.

More Species, More Hybrids, and More Alleles

The model we introduced above consists of a single-hybrid phylogenetic network on four species and assumes that a single allele is sampled per species. This model was assumed in the inference algorithm, as well as the simulation study. If we still assume a single hybrid node and a single allele per species, then considering more than four species may have an effect on the model and the performance of the inference algorithm. If the number of species under the hybrid node remains 2, then the probabilities of inheritance of the alleles in the ancestral hybrid population from A or D remain unchanged. However, if the number of species under the hybrid node increases beyond 2, then the ancestral

hybrid population may have more than 2 alleles, and the inheritance probabilities from each of the 2 parents (A and D) follow a binomial distribution.

A similar scenario occurs if we keep the topology of the network in Figure 1 unchanged; yet, we allow multiple alleles to be sampled. In this case, the time t_3 in Figure 1 determines the number of alleles entering (looking backward in time) the ancestral hybrid population, and the number of alleles inherited from A is binomially distributed. Furthermore, in this case, the model has to incorporate the general formula for the probability of m alleles coalescing into k alleles within time t, $g_{m,k}(t)$.

In the case when more than a single hybrid node is present in the phylogenetic network, two issues need to be addressed: 1) efficient search of the network space, and 2) efficient calculation of valid coalescent histories within the branches of a given network, as well as computing the number of extra lineages on a network. For a network with multiple hybrids, multiple paths through the network can result in identical valid coalescent histories (e.g., the valid coalescent history where all coalescence events occur above the root). An important question to address here is whether all such paths should contribute to the probability of observing a gene tree topology, or only a single one of them should contribute. Furthermore, because the concept of *parental species trees* does not apply here, an efficient procedure for computing the probabilities of all gene tree topologies is needed.

The bottom line is that developing methods for larger data sets (more species, more hybrids, or more alleles) would require major extensions of the model that is considered here, as well as the models considered for other methods available in the literature.

FUNDING

This work was supported in part by National Science Foundation grant CCF-0622037, R01LM009494 from the National Library of Medicine, and an Alfred P. Sloan Research Fellowship to L.N. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the NSF, National Library of Medicine, the National Institutes of Health, or the Alfred P. Sloan Foundation. J.H.D. was funded by the New Zealand Marsden Fund.

ACKNOWLEDGMENTS

Jack Sullivan, Laura S. Kubatko, and two anonymous reviewers provided comments on an earlier version of this manuscript that significantly improved it in terms of technical contents as well as presentation.

REFERENCES

Arnold M.L. 1997. Natural hybridization and evolution. Oxford: Oxford University Press.

Bloomquist E.W., Suchard M.A. 2010. Unifying vertical and nonvertical evolution: a stochastic ARG-based framework. Syst. Biol. 59: 27–41.

Cranston, K.A., Hurwitz B., Ware D., Stein L., Wing R.A. 2009. Species trees from highly incongruent gene trees in rice. Syst. Biol. 58: 489–500.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Degnan J.H., Salter L.A. 2005. Gene tree distributions under the coalescent process. Evolution. 59:24–37.

Dunn B., Sherlock G. 2008. Reconstruction of the genome origins and evolution of the hybrid lager yeast *Saccharomyces pastorianus*. Genome Res. 18:1610–1623.

Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution. 63:1–19.

Edwards S.V., Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. U.S.A. 104:5936–5941.

Gonzalez, S.S., Barrio E., Querol A. 2008. Molecular characterization of new natural hybrids of *Saccharomyces cerevisiae* and *S. kudriavzevii* in brewing. Appl. Environ. Microbiol. 74:2314–2320.

Gonzalez S.S., Gallo L., Climent M.A., Barrio E., Querol A. 2007. Ecological characterization of natural hybrids from *Saccharomyces cerevisiae* and *S. kudriavzevii*. Int. J. Food Microbiol. 116:11–18.

Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.

Holland B., Benthin S., Lockhart P., Moulton V., Huber K. 2008. Using supernetworks to distinguish hybridization from lineage-sorting. BMC Evol. Biol. 8:202.

Jin G., Nakhleh L., Snir S., Tuller T. 2007. Inferring phylogenetic networks by the maximum parsimony criterion: a case study. Mol. Biol. Evol. 24:324–337.

Joly S., McLenachan P.A., Lockhart P.J. 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. Am. Nat. 174:E54–E70.

Jukes T., Cantor C. 1969. Evolution of protein molecules. In: Munro H., editor. Mammalian Protein Metabolism. New York: Academic Press. p. 21–132

Kubatko L.S. 2009. Identifying hybridization events in the presence of coalescence via model selection. Syst. Biol. 58:478–488.

Kubatko L.S., Carstens B.C., Knowles L.L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics. 25:971–973.

Kuo C.-H., Wares J.P., Kissinger J.C. 2008. The Apicomplexan whole-genome phylogeny: an analysis of incongurence among gene trees. Mol. Biol. Evol. 25:2689–2698.

Linder C.R., Rieseberg L.H. 2004. Reconstructing patterns of reticulate evolution in plants. Am. J. Bot. 91:1700–1708.

Liu L., Pearl D.K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.

Liu L., Yu L., Kubatko L.S., Pearl D.K., Edwards S.V. 2009. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53:320–328.

Maddison W. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.

Mallet J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20:229–237.

Mallet J. 2007. Hybrid speciation. Nature. 446:279–283.

Meng C., Kubatko L.S. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theor. Popul. Biol. 75:35–45.

Nakhleh L. 2010. Evolutionary phylogenetic networks: models and issues. In: Heath L., Ramakrishnan N., editors. The problem solving handbook for computational biology and bioinformatics. New York: Springer. p. 125–158.

Nordborg M. 1998. On the probability of neanderthal ancestry. Am. J. Hum. Genet. 63:1237–1240.

Pollard D.A., Iyer V.N., Moses A.M., Eisen M.B. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. PLoS Genet. 2:e173.

Rambaut A., Grassly N.C. 1997. Seq-gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comp. Appl. Biosci. 13:235–238.

Rannala B., Yang Z. 2008. Phylogenetic inference using whole genomes. Annu. Rev. Genomics Hum. Genet. 9:217–231.

Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425:798–804.

Syring J., Willyard A., Cronn R., Liston A. 2005. Evolutionary relationships among *Pinus* (Pinaceae) subsections inferred from multiple low-copy nuclear loci. Am. J. Bot. 92:2086–2100.

Tavaré S. 1984. Lines-of-descent and genealogical processes, and their application in population genetic models. Theor. Popul. Biol. 26:119–164.

Than C., Nakhleh L. 2009. Species tree inference by minimizing deep coalescences. PLoS Comput. Biol. 5:e1000501.

Than C., Ruths D., Innan H., Nakhleh L. 2007. Confounding factors in HGT detection: Statistical error, coalescent effects, and multiple solutions. J. Comput. Biol. 14:517–535.

Than C., Ruths D., Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics 9:322.

Than C., Sugino R., Innan H., Nakhleh L. 2008. Efficient inference of bacterial strain trees from genome-scale multi-locus data. Bioinformatics 24:i123–i131.