Theoretical Population Biology 75 (2009) 35–45

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model

Chen Meng [a], Laura Salter Kubatko [b,*]

[a] Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131, USA

[b] Departments of Statistics and Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210, USA

ARTICLE INFO

Article history:
Received 11 March 2008
Available online 5 November 2008

Keywords:
Phylogenetics
Hybridization
Introgression
Hybrid speciation
Incomplete lineage sorting
Coalescence

ABSTRACT

The application of phylogenetic inference methods, to data for a set of independent genes sampled randomly throughout the genome, often results in substantial incongruence in the single-gene phylogenetic estimates. Among the processes known to produce discord between single-gene phylogenies, two of the best studied in a phylogenetic context are hybridization and incomplete lineage sorting. Much recent attention has focused on the development of methods for estimating species phylogenies in the presence of incomplete lineage sorting, but phylogenetic models that allow for hybridization have been more limited. Here we propose a model that allows incongruence in single-gene phylogenies to be due to both hybridization and incomplete lineage sorting, with the goal of determining the contribution of hybridization to observed gene tree incongruence in the presence of incomplete lineage sorting. Using our model, we propose methods for estimating the extent of the role of hybridization in both a likelihood and a Bayesian framework. The performance of our methods is examined using both simulated and empirical data.

1. Introduction

Evolutionary inference via phylogenetic trees has been increasingly performed on DNA sequence data from multiple genes for collections of species. A crucial issue that arises when multi-gene data are analyzed is that each gene has its own evolutionary history, which may or may not be congruent with the evolutionary history of the species as a whole (Tateno et al., 1982; Nei, 1987). While this issue has long been recognized (Maddison, 1997; Pamilo and Nei, 1988), methods to appropriately model variation in evolutionary histories in a phylogenetic context have only been developed recently, and such methods usually consider only one of the possible evolutionary processes that lead to variability in the histories of individual genes (Maddison and Knowles, 2006; Edwards et al., 2007).

Possible evolutionary processes leading to incongruence between gene and species phylogenies include hybridization (Arnold, 1997; Mallet, 2005, 2007), horizontal gene transfer (Medigue et al., 1991; Valdez and Pinero, 1992), incomplete lineage sorting, and gene duplication and extinction (Maddison, 1997; Guigo et al., 1996). Incomplete lineage sorting, also called deep coalescence, is the failure of ancestral gene copies to coalesce into a common ancestral copy until earlier than the previous speciation event, and is perhaps the best-studied of these biological mechanisms. Incomplete lineage sorting is mathematically modeled by the coalescent process (Kingman, 1982; Hudson, 1983; Tajima, 1983), which can be used to compute the probabilities of individual gene tree topologies for a given species phylogeny when that phylogeny represents the historical relationships among these species (or populations) (Tajima, 1983; Takahata and Nei, 1985; Pamilo and Nei, 1988; Rosenberg, 2002; Degnan and Salter, 2005). Rannala and Yang (2003) also used this model to derive the joint density for gene tree topology and branch lengths given a species tree. These developments have led to several techniques which specifically incorporate the coalescent in multi-gene phylogenetic analysis (Maddison and Knowles, 2006; Liu and Pearl, 2007; Edwards et al., 2007; Than et al., 2007; Mossel and Roch, 2008).

Hybridization is another well-documented evolutionary process that leads to incongruence in gene and species topologies. Hybridization is an important evolutionary mechanism in plants and bacteria and has been increasingly documented in animals (Rieseberg, 1997; Dowling and Secor, 1997; Arnold, 1997; Baack and Rieseberg, 2007). It has been estimated that at least 25% of plant species and 10% animal species hybridize (Mallet, 2005, 2007). Inclusion of taxa that have resulted from hybridization leads to problems in phylogenetic tree reconstruction as such taxa do not meet the usual assumption of evolution from a common ancestor through a bifurcating process (Hennig, 1966).

Some efforts have been made to resolve and represent a phylogenetic network that allows the occurrence of historical

* Corresponding author.
E-mail address: lkubatko@stat.osu.edu (L.S. Kubatko).

0040-5809/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2008.10.004

hybridization. For example, some researchers have chosen to exclude hybrids from the analysis (Posada, 2002). They identify hybrids before the phylogenetic analysis, remove the hybrids prior to the analysis, and then connect them to the putative parents on the inferred tree (Posada, 2002). Because hybrids generally maintain morphological or molecular intermediacy and/or combination between the parents, observed intermediacy/combination is often used to identify a hybrid and its parents in this approach. Other researchers have developed different methods to detect and reconstruct hybridization from the joint analysis of hybrids and putative parental taxa. For example, Rieseberg and Morefield (1995) developed a method to identify a hybrid according to the amount of homoplasy which reflects the degree of intermediacy between the hybrid and its parents. Sang and Zhong (2000) proposed a test to distinguish between hybridization and lineage sorting based on differences in the divergence times in incongruent gene trees. Gauthier and Lapointe (2007) presented a Hybrid Detection Criterion, along with a statistical test that assesses whether a putative hybrid is consistently intermediate between its postulated parents, with respect to the other taxa. Finally, Than et al. (2007) developed computational methods to infer a tree(s) that minimizes the number of postulated horizontal transfer events required to explain the data.

Methods to model reticulate evolution have also led to several useful techniques to infer phylogenetic relationships that are not constrained by a typical bifurcating tree. For example, Strimmer and Moulton (2000) presented a framework for inferring a maximum-likelihood phylogenetic network from a set of DNA sequences based on directed graphical models. They converted any given phylogenetic network representing an evolutionary hypothesis into a directed graphical model and derived the joint probability distribution involving all nodes of the network dependent on branch lengths and recombination parameters. Their work enables evaluation of phylogenetic networks on a statistically sound basis. Algorithms for constructing phylogenetic networks have also been developed on the basis of distance matrices. Xu (2000) conducted phylogenetic reconstruction under the pure drift model and the mutation model by a least-squares method using genetic distances between taxa for gene frequency data. Legendre and Makarenkov (2002) developed a method for reconstructing reticulation networks from empirical distance matrices.

As our sophistication in modeling complex evolutionary processes improves, methods which incorporate several of these processes simultaneously should become possible. The method developed by Sang and Zhong (2000) did allow both hybridization and deep coalescence, but their method was later criticized (Holder et al., 2001) because they failed to assume any variance in coalescent time except in cases of lineage sorting, and thus their test was shown not powerful enough to discriminate between hybridization and lineage sorting in data with realistic levels of variability. Buckley et al. (2006) used simulation to examine whether the extent of observed incongruence in the placement of a putative hybrid taxon in their empirical data was consistent with what might be expected under the coalescent model. Than et al. (2007) proposed a model that incorporates both horizontal gene transfer and incomplete lineage sorting which shares some similarity to the model we propose below. In particular, they demonstrated how to compute the probabilities associated with various gene tree topologies in the case of three taxa when both horizontal gene transfer and incomplete lineage sorting can lead to variation in observed gene trees. While these methods are useful starting points in modeling hybridization in the coalescent framework, methods to detect the level or probability of hybridization between species in this context are still in their infancy, as noted by Mallet (2005). In this paper, we propose a model for examining both processes in a phylogenetic context, with the explicit goal of testing for hybridization.

In our model, we define hybridization as gene flow between distinct species. When repeated back-crossing occurs, the genetic material of one species may become integrated into another, a process known as introgression (Baack and Rieseberg, 2007). This introgression may eventually lead to speciation (Baack and Rieseberg, 2007), and is thus a mechanism for hybrid speciation. We further assume independent evolution in a sample of genes that allows incongruence among a sample of gene tree topologies to be due to both deep coalescence and hybrid speciation. We use both maximum likelihood and Bayesian approaches to distinguish hybridization in the presence of coalescence from coalescence alone as the source of conflict in individual gene histories. In the likelihood framework, we derive asymptotic results for the distribution of our statistic, and examine the small-sample behavior using simulation. In the Bayesian framework, inference is performed through implementation of a Markov chain Monte Carlo (MCMC) algorithm. This method has the advantage of allowing existing knowledge concerning hybrid speciation among the taxa of interest to be incorporated via prior distributions. Finally, we illustrate how our method can be used to assess hybridization with both empirical and simulated data.

2. Materials and methods

2.1. A model for hybridization with coalescence

Before introducing our model, we begin with a brief review of the coalescent process as it relates to the distribution of gene trees, given a species phylogeny. The basic coalescent model makes the assumptions that populations are large and panmictic, that the mutation process is selectively neutral, and that population sizes are constant within populations (Kingman, 1982; Nordborg, 2001). When incorporating the coalescent into a phylogenetic framework, it is also standard to make assumptions of no recombination within genes and no migration or other horizontal gene transfer across the species-level phylogeny. The coalescent traces the history of lineages back in time, where time is measured in number of generations (Kingman, 1982; Hudson, 1983; Tajima, 1983). Assuming a population of $2N_e$ alleles, the probability that a pair of alleles share an ancestor in the previous generation is $\frac{1}{2N_e}$. The number of generations until a pair of alleles share an ancestor then follows a geometric distribution with parameter $\frac{1}{2N_e}$ under the assumption of constant population size. This is well-approximated by an exponential distribution when the number of generations becomes large (Kingman, 1982). Most calculations carried out under the coalescent model use this continuous-time approximation.

When viewed in a phylogenetic setting, we consider a fixed species tree (a tree which gives the true sequence of speciation events) with branch lengths given in coalescent units of $s/(2N_e)$, where s is the number of generations and N_e is the effective population size (number of diploid individuals). The basic coalescent model described above then operates independently within each branch of the species tree. Fig. 1 illustrates this process by describing the possible relationships between gene trees and species trees under the coalescent process for the simple case of three taxa. In each subfigure, the bolder, outlined tree gives the topology of the species-level relationships assumed here, while the thin lines drawn inside of the species tree give gene tree relationships. The upside down v's in the species tree represent speciation events. Coalescent events on the gene trees are marked by dots.

Note that for the example in Fig. 1, there are two possible scenarios, called coalescent histories, which lead to congruence (agreement) between the gene tree and the species tree: taxa B and C can coalesce in the interval between the speciation

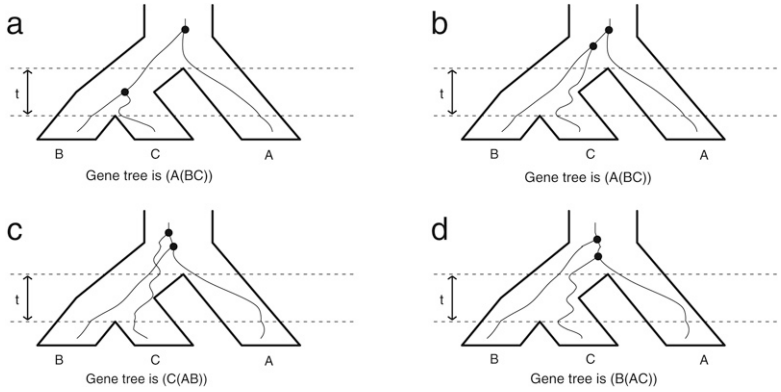

Fig. 1. The coalescent model for the relationship between gene trees and species trees in the case of three taxa is shown. Each subfigure shows the assumed species tree (bold lines) with a particular gene history drawn inside the species tree (thin lines). Speciation events are represented by dotted lines and upside down v's in the species tree, while gene divergence (coalescent) events are marked by dots in the gene trees. Note that more than one history may correspond to a particular gene tree topology (e.g., the histories (a) and (b) have the same gene tree topology).

event splitting populations A and B–C and the speciation event which divides populations B and C (Fig. 1(a)), or taxa B and C can coalesce prior (looking backward in time) to the speciation event splitting populations A and B–C (Fig. 1(b)). Because times to coalescent events are exponentially distributed, the probability of the history shown in Fig. 1(a) is just the probability that the time to coalescence of lineages B and C is less than t, which is $1 - e^{-t}$, where t is the length of the interval between the two speciation events, measured in coalescent units. When lineages B and C do not coalesce in the interval between speciation events, which happens with probability e^{-t}, then all three lineages are available to coalesce prior to the first speciation event. Assuming that each pair is equally likely to be the first to coalesce, the probability associated with each gene tree is simply $\frac{1}{3}e^{-t}$ (Fig. 1(b)–(d)). Note that the probability of congruence between the gene and species tree depends directly on t, the branch length in the species tree. As t increases, the chance of coalescence in the interval between speciation events increases, and the probability of observing an incongruent gene tree decreases. When t is short, it is difficult for lineages B and C to coalesce in this interval, and incongruence is more probable. We further note that, since t is measured in units of $s/(2N_e)$, a small value of t can result from either a small number of generations between divergence events, or a large effective population size.

Computation of gene tree probabilities is deceptively simple in the three-taxon case because the number of histories corresponding to each gene tree topology is limited (either one, for the gene trees in Fig. 1(c) and (d), or two, for the gene tree common to Fig. 1(a) and (b)). As the number of taxa increases, however, each possible gene tree topology will have many histories that will need to be considered in computing this probability. Although the relationships between gene trees and species trees under the coalescent have been studied for some time (Tavaré, 1984; Takahata and Nei, 1985; Pamilo and Nei, 1988; Takahata, 1989; Rosenberg, 2002), computation of the gene tree probabilities for arbitrary species trees was previously limited to five or fewer taxa (Pamilo and Nei, 1988; Rosenberg, 2002). However, a recently developed algorithm (Degnan and Salter, 2005) allows the computation of the entire probability distribution of gene tree topologies. We use this probability distribution directly in our model.

We now give an extension of the basic model described above to incorporate hybridization. As previously done (Rieseberg and Wendel, 1993; Rieseberg, 1997), we assume that speciation by hybridization results in a mosaic genome, in which a randomly selected gene will be derived from one of the two parental species. This means that when a randomly selected gene is considered in a phylogenetic analysis, it will have descended from one of two

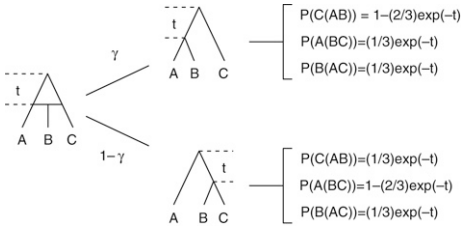

Fig. 2. The hybrid speciation model proposed here is shown. The leftmost panel depicts the true phylogenetic relationships between taxa A, B, and C, with hybridization represented by a horizontal line. The middle panel shows the two topologies formed by selecting either the A or the C lineage to be the "parent" of B for a given gene, which occur with probability γ and $1 - \gamma$, respectively. The rightmost panel gives the probabilities of the three possible gene trees under the two scenarios (Pamilo and Nei, 1988). t represents the time between speciation events on the parental trees.

possible parental trees. To make the model more concrete, consider the three-taxon tree in Fig. 2. The species tree in the leftmost panel of the figure represents the true sequence of speciation events, where the horizontal line indicates the formation of hybrid species B from parental species A and C. Our model supposes that when a single gene from the hybrid species is selected for analysis, its most recent common ancestor occurs with taxon A (resulting in the tree in the top of the middle panel in Fig. 2) with probability γ and with taxon C (resulting in the tree in the bottom of the middle panel in Fig. 2) with probability $1 - \gamma$. Once the tree for that particular gene has been selected, the coalescent process operates for that gene, using the selected tree as the species tree, and probabilities for particular gene trees can be derived (Pamilo and Nei, 1988) (see rightmost panel in Fig. 2). The model is analogous for larger trees, in the sense that any hypothesized hybridization event leads to two possible species-level relationships, which can then be used to compute the required gene trees distributions using the techniques of Degnan and Salter (2005).

Because γ is the parameter associated with hybridization, estimation or tests concerning past hybridization can be carried out via inference about γ. This parameter can also be given a more direct biological interpretation as the proportion of the genome arising from a particular parent (taxon A in the example in Fig. 2). A special case is that γ is equal to zero or one, which indicates that no hybridization has occurred. We propose two methods for estimation of γ, both based on the likelihood function. In both cases, we assume that the data available consist of a sample of gene tree topologies which are assumed to be known without error. Specifically, let $Data = \{g_1, g_2, \ldots, g_N\}$, where g_i represents the topology for the ith gene, with $i = 1, 2, \ldots, N$ for a total of N genes. We assume that genes are sampled in such a way that,

conditional on the species tree, their topologies are independent and follow the hybridization model described above. Note that only information on gene tree topologies is used in the model; branch length information is not incorporated. However, species tree branch lengths (determined by speciation times) are needed to compute the probability distribution on gene tree topologies. For example, in Figs. 1 and 2, the probability associated with each gene tree topology is a function of t, the time between speciation events in the parental trees.

2.2. Maximum likelihood estimation

Under the assumption of an independent and identically distributed (i.i.d.) sample of gene trees from the hybridization model, the likelihood function for a given species tree with a specified location for the hybridization event is the product of probabilities of all observed gene trees under our proposed model,

$$\begin{aligned} L(\gamma, \mathbf{t}|Data) &= \prod_{i=1}^{N} P(g_i|\gamma, \mathbf{t}) \\ &= \prod_{i=1}^{N} \{\gamma P(g_i|\tau_1, \mathbf{t}) + (1-\gamma)P(g_i|\tau_2, \mathbf{t})\} \end{aligned} \quad (1)$$

where τ_1 and τ_2 are the two possible parental topologies and $\mathbf{t}$ is a vector of species tree branch lengths. $P(g_i|\tau_1, \mathbf{t})$ and $P(g_i|\tau_2, \mathbf{t})$ can be computed using the COAL program (Degnan and Salter, 2005). In the general case of a species tree with n taxa, the likelihood will be a function of $n-1$ parameters: $n-2$ species tree branch lengths and the γ parameter. In this case, the number of possible gene trees will be $M = \frac{(2n-3)!}{2^{n-2}(n-2)!}$, and letting m_j be the number of times tree j ($j \in \{1, 2, \ldots M\}$) is observed in the data, we can write the log likelihood as

$$\begin{aligned} \log L(\gamma, \mathbf{t}|Data) &= \log\left(\prod_{j=1}^{M} P(g_j|\gamma, \mathbf{t})^{m_j}\right) \\ &= \sum_{j=1}^{M} m_j \log P(g_j|\gamma, \mathbf{t}). \end{aligned} \quad (2)$$

The joint maximum likelihood estimates (MLEs) of γ and the branch lengths $\mathbf{t}$ are given by the values that maximize this likelihood function. Because calculation of the likelihood function can be performed rapidly, these joint MLEs can be simply computed using a grid search of the plausible parameter space when n is small. We implement this method in the case $n = 4$ when we test our method below. Extensions to larger trees are described in Section 5.

We next consider estimation of the variance in the parameter estimates and derivation of asymptotic distributional results. Note that when there is no hybridization, the true value of γ may be on the boundary of the parameter space [0, 1]. Self and Liang (1987) have derived the asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions such as this. Applying their results to our model, it is possible to show that the joint MLEs of γ and $\mathbf{t}$, denoted by the parameter vector $\hat{\Theta}_{MLE} = (\hat{\gamma}_{MLE}, \hat{\mathbf{t}}_{MLE})$, are statistically consistent as the number of genes, N, goes to ∞ (we have checked this for $n = 3$; the extension to other cases is straightforward). When it can be additionally assumed that γ is in the open interval (0, 1) (i.e, that there has been at least some gene flow between the parental species), then $\hat{\Theta}_{MLE}$ is also asymptotically efficient, i.e.,

$$\sqrt{N}(\hat{\Theta}_{MLE} - \Theta) \longrightarrow Normal(0, I^{-1}(\Theta)) \quad (3)$$

where $I(\Theta)$ is the Fisher information matrix.

To make this result more concrete, consider the case of three taxa (Fig. 2). Then there is a single species tree branch length, which we denote by t. The Fisher information matrix, $I(\Theta)$, is given by

$$\begin{pmatrix} -\sum_{j=1}^{M} Nf_j(\gamma, t)\left(\frac{\partial^2}{\partial\gamma^2}\log(f_j(\gamma, t))\right) & -\sum_{j=1}^{M} Nf_j(\gamma, t)\left(\frac{\partial^2}{\partial\gamma\,\partial t}\log(f_j(\gamma, t))\right) \\ -\sum_{j=1}^{M} Nf_j(\gamma, t)\left(\frac{\partial^2}{\partial t\,\partial\gamma}\log(f_j(\gamma, t))\right) & -\sum_{j=1}^{M} Nf_j(\gamma, t)\left(\frac{\partial^2}{\partial t^2}\log(f_j(\gamma, t))\right) \end{pmatrix} \quad (4)$$

where $f_j(\gamma, t)$ is the likelihood of the data for tree j, i.e., $f_j(\gamma, t) = P(g_j|\gamma, t)$.

We obtain estimated variances of our parameter estimates based on the observed Fisher information by evaluating the entries in the above matrix for the sample of gene trees observed in our data. Thus the estimated variances would be

$$\hat{var}(\hat{\gamma}) = \frac{1}{D}\sum_{j=1}^{M} Nf_j(\gamma, t)\left(\frac{\partial^2}{\partial t^2}\log f_j(\gamma, t)\right) \quad (5)$$

and

$$\hat{var}(\hat{t}) = \frac{1}{D}\sum_{j=1}^{M} Nf_j(\gamma, t)\left(\frac{\partial^2}{\partial \gamma^2}\log f_j(\gamma, t)\right) \quad (6)$$

where D is the determinant of the matrix in (4),

$$\begin{aligned} D = &\left(\sum_{j=1}^{M} Nf_j(\gamma, t)\left(\frac{\partial^2}{\partial\gamma^2}\log f_j(\gamma, t)\right)\right) \\ &\times \left(\sum_{j=1}^{M} Nf_j(\gamma, t)\left(\frac{\partial^2}{\partial t^2}\log f_j(\gamma, t)\right)\right) \\ &- \left(\sum_{j=1}^{M} Nf_j(\gamma, t)\left(\frac{\partial^2}{\partial t\,\partial\gamma}\log\left(f_j(\gamma, t)\right)\right)\right)^2. \end{aligned} \quad (7)$$

The entries in Eqs. (5)–(7) are evaluated at the MLEs of γ and t in computing the estimated variances.

The asymptotic results above are based on the assumption that hybridization occurs to some extent, so that $\gamma \in (0, 1)$. While our real interest will often be testing whether hybridization has occurred (i.e., whether $\gamma = 0$ or $\gamma = 1$), the results above are presented to motivate our methodology for estimating variances. When the true value of γ is equal to 0 or 1, it is still possible to obtain an expression for the asymptotic distribution of $\hat{\Theta}$ using the results of Self and Liang (1987), however this distribution is a more complicated mixture distribution. In this case, the estimate obtained using the observed Fisher information is expected to overestimate the variances and is therefore conservative when $\gamma = 0$ or $\gamma = 1$.

Maximum likelihood estimation of γ also enables construction of a likelihood ratio test, which can be used to assess the evidence for hybridization. To test whether hybridization has played a role in speciation, we can use the likelihood ratio statistic, Δ, for testing $H_0 : \gamma = 0$ vs. $H_a : \gamma \neq 0$,

$$\Delta = -2\ln\frac{L(\gamma_0 = 0, \hat{\mathbf{t}}_0|Data)}{L(\hat{\gamma}_{MLE}, \hat{\mathbf{t}}_{MLE}|Data)} \quad (8)$$

where $\hat{\mathbf{t}}_0$ is the MLE of the branch length vector $\mathbf{t}$ when $\gamma = 0$. The test is carried out by comparing Δ to a 50:50 mixture of χ_1^2 and a point mass at 0 (Self and Liang, 1987).

2.3. Bayesian estimation

Because the maximum likelihood results presented above are based on large sample results, we expect that this method will perform best for data sets in which a large number of genes have been sampled. However, for many problems, information is available for only a handful of genes, and so we develop a Bayesian approach as well. The Bayesian approach has the additional advantage of allowing the investigator to incorporate any relevant

prior knowledge concerning the likelihood of hybridization (e.g., do the two species under consideration hybridize in the lab?) into the estimation procedure.

We begin by specifying a mixed Beta distribution as the prior for γ, which is a logical choice since γ is between 0 and 1. In addition, the two parameters of the Beta distribution allow for a great deal of flexibility in the prior, with the uniform distribution on the interval (0, 1) as a special case. A hyperparameter ϵ is then introduced to reflect the overall prior belief concerning the likelihood of hybridization between two taxa. We assume that ϵ follows a Bernoulli distribution with parameter θ,

$$P(\epsilon|\theta) = \theta^{\epsilon}(1-\theta)^{1-\epsilon}. \tag{9}$$

The value of θ is determined based on the investigator's prior belief concerning the existence of hybridization for the taxa under consideration. When ϵ is 1, the prior for γ follows a Beta distribution with parameters α_1 and β_1, which are chosen so that the prior distribution is concentrated at 0. When ϵ is 0, the prior for γ is another Beta distribution with parameters α_2 and β_2, which are chosen to reflect the range and likely values of γ. The mixed prior for γ is

$$P(\gamma|\epsilon) = \epsilon\frac{\gamma^{\alpha_1-1}(1-\gamma)^{\beta_1-1}}{B(\alpha_1,\beta_1)} + (1-\epsilon)\frac{\gamma^{\alpha_2-1}(1-\gamma)^{\beta_2-1}}{B(\alpha_2,\beta_2)}. \tag{10}$$

This hierarchical set-up is modeled after the technique used by George and McCulloch (1993) in problems of variable selection. The intuition behind the model is that θ represents the prior probability that hybridization does not occur at all, and therefore the posterior probability that $\epsilon = 1$ will provide information concerning evidence for hybridization in the data. The smaller the probability that $\epsilon = 1$, the stronger the evidence.

The probability of the data given the parameters (i.e., the likelihood) is given by Eq. (1), except that here we assume that γ is a random variable with prior distribution (10). In addition to γ, the species tree branch length parameters $\mathbf{t}$ are also unknown. While a Bayesian procedure could be developed to incorporate uncertainty in these parameters as well, we instead adopt a pseudo-empirical Bayesian approach here and set these parameters to their MLEs. Reference to these parameters is therefore omitted in the formulation below. Next we note that since $P(g_i|\tau_1)$ and $P(g_i|\tau_2)$ can be computed using the method of Degnan and Salter (2005), the likelihood is in fact a polynomial function of γ and can be written as

$$P(Data|\gamma,\epsilon) = \sum_{j=0}^{N} c_j\gamma^j \tag{11}$$

where the c_j are determined by $P(g_i|\tau_1)$ and $P(g_i|\tau_2)$ for $i = 1, 2, \ldots, N$ (see the Appendix for details). Note that this representation of the likelihood simplifies derivation of the conditional distributions in Eqs. (13) and (15) below. The joint distribution is

$$P(Data, \gamma, \epsilon) = P(Data|\gamma)P(\gamma|\epsilon)P(\epsilon). \tag{12}$$

Our interest is in estimation of the posterior distribution of the parameters ϵ and γ. Because this cannot be computed analytically, we use Markov chain Monte Carlo (MCMC) to approximate the posterior distribution. In particular, a Gibbs sampling method is appropriate here, because the joint posterior distribution of γ and ϵ is intractable, but the conditional distributions can be easily derived. For example, the full conditional distribution of γ given the data and ϵ is

$$P(\gamma|Data,\epsilon) = \frac{P(Data,\gamma,\epsilon)}{P(Data,\epsilon)} \tag{13}$$
$$= \frac{\left\{\sum_{j=0}^{N} c_j\gamma^j\right\}\left\{\epsilon\frac{\gamma^{\alpha_1-1}(1-\gamma)^{\beta_1-1}}{B(\alpha_1,\beta_1)} + (1-\epsilon)\frac{\gamma^{\alpha_2-1}(1-\gamma)^{\beta_2-1}}{B(\alpha_2,\beta_2)}\right\}\theta^{\epsilon}(1-\theta)^{1-\epsilon}}{\sum_{j=0}^{N}\frac{c_j\epsilon\theta^{\epsilon}(1-\theta)^{1-\epsilon}B(\alpha_1+j,\beta_1)}{B(\alpha_1,\beta_1)} + \sum_{j=0}^{N}\frac{c_j(1-\epsilon)\theta^{\epsilon}(1-\theta)^{1-\epsilon}B(\alpha_2+j,\beta_2)}{B(\alpha_2,\beta_2)}}$$

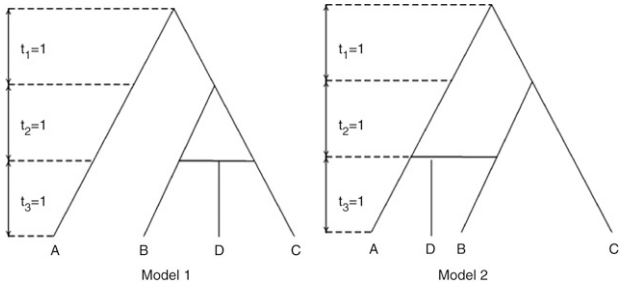

Fig. 3. Hybrid speciation models used for simulation study. Branch lengths are given in coalescent units (number of $2N_e$ generations, where N_e is the effective population size). In Model 1, hybridization occurs between sister taxa B and C to form taxon D, while in Model 2, non-sister taxa A and B lead to formation of D. Because A and B are less closely related, the role of hybridization in the formation of taxon D should be easier to detect assuming Model 2 than assuming Model 1.

where $B(\alpha_i, \beta_j)$ is given by

$$B(\alpha_i, \beta_j) = \frac{\Gamma(\alpha_i)\Gamma(\beta_j)}{\Gamma(\alpha_i+\beta_j)}, \tag{14}$$

and the full conditional distribution of ϵ given the data and γ is

$$P(\epsilon|Data,\gamma) = \frac{P(Data,\gamma,\epsilon)}{P(Data,\gamma)} \tag{15}$$
$$= \frac{\left\{\sum_{j=0}^{N} c_j\gamma^j\right\}\left\{\epsilon\frac{\gamma^{\alpha_1-1}(1-\gamma)^{\beta_1-1}}{B(\alpha_1,\beta_1)} + (1-\epsilon)\frac{\gamma^{\alpha_2-1}(1-\gamma)^{\beta_2-1}}{B(\alpha_2,\beta_2)}\right\}\theta^{\epsilon}(1-\theta)^{1-\epsilon}}{\left\{\sum_{j=0}^{N} c_j\gamma^j\right\}\frac{\gamma^{\alpha_1-1}(1-\gamma)^{\beta_1-1}}{B(\alpha_1,\beta_1)}\theta + \left\{\sum_{j=0}^{N} c_j\gamma^j\right\}\frac{\gamma^{\alpha_2-1}(1-\gamma)^{\beta_2-1}}{B(\alpha_2,\beta_2)}(1-\theta)}.$$

Although the conditional distributions in Eqs. (13) and (15) are easily written down, we have chosen to use a Metropolis within Gibbs algorithm at each iteration of the chain, instead of directly sampling from these full distributions. Thus at each iteration of the algorithm, we simulate a value from $P(\gamma|Data,\epsilon)$ using Metropolis within Gibbs, followed by a value from $P(\epsilon|Data,\gamma)$ using Metropolis within Gibbs. Sample collection starts after a certain number of iterations (the burnin period), and every kth iterate following the burnin is recorded. Convergence of the chain to the posterior distribution of interest is assessed by checking the trace plots and also by comparing the results of simultaneous multiple chains with different random starting points, using standard diagnostics such as the Gelman–Rubin Potential Scale Reduction Factor (PSRF) (Gelman and Rubin, 1992).

2.4. Simulations

We used simulation studies to evaluate the estimates of γ obtained by the maximum likelihood and Bayesian methods for a four-taxon tree. Assuming that a hybridization event occurs, there are two possible sequences of speciation events, which are depicted in Fig. 3. The left panel of Fig. 3 gives the sequence of speciation events in which the hybrid taxon D diverged from two species, B and C, which have the same most recent common ancestor. Under this model, the probability that a randomly selected gene of D is inherited from B is γ and the probability that the gene is derived from C is $1-\gamma$. The right panel gives the other possibility, with D being the hybrid of A and B, which do not share the same most recent common ancestor. Under this model, A is the parent of D for a given gene with probability γ and B is the parent of D for a given gene with probability $1-\gamma$.

For each hybridization model, we considered six values for γ: 0, 0.1, 0.2, 0.3, 0.4, and 0.5. For each value of γ, we considered three values for the number of gene trees in each sample. The selected sample sizes were $N = 10$, 20, and 50 for the maximum

Table 1
Results of the maximum likelihood analysis of the simulated data for both models. The left two columns give the true value of γ used to generate the data and the number of gene trees simulated with COAL (i.e., the sample size). The remaining columns give, for each model, the mean of the MLEs of γ over the 100 simulated data sets, the variance in these 100 MLEs, the mean of the estimated variances using our estimator of the observed Fisher information (±standard deviation), and the estimated power of the likelihood ratio test (i.e., the proportion of times the null hypothesis $\gamma = 0$ was rejected in the 100 trials).

True γ	Sample size	Model 1				Model 2			
		Average MLE	Observed variance	Estimated variance	Estimated power	Average MLE	Observed variance	Estimated variance	Estimated power
0.5	10	0.4820	0.0889	0.6674 ± 3.4004	0.41	0.4884	0.0257	0.0280 ± 0.0063	0.98
0.4	10	0.3619	0.0597	0.6109 ± 3.4276	0.31	0.4165	0.0292	0.0274 ± 0.0080	0.90
0.3	10	0.3218	0.0475	0.4748 ± 2.6929	0.34	0.3167	0.0232	0.0238 ± 0.0088	0.80
0.2	10	0.2058	0.0338	0.1947 ± 0.7476	0.21	0.2311	0.0176	0.0203 ± 0.0084	0.68
0.1	10	0.1577	0.0284	0.2688 ± 1.0122	0.19	0.1167	0.0097	0.0134 ± 0.0079	0.31
0	10	0.0728	0.0082	0.0613 ± 0.1116	0.09	0.0247	0.0012	0.0049 ± 0.0038	0.06
0.5	20	0.4855	0.0285	0.0452 ± 0.0881	0.67	0.4912	0.0196	0.0141 ± 0.0023	1.00
0.4	20	0.3747	0.0214	0.0728 ± 0.3109	0.46	0.3939	0.0161	0.0141 ± 0.0025	0.96
0.3	20	0.3084	0.0282	0.0987 ± 0.5780	0.44	0.2874	0.0108	0.0123 ± 0.0026	0.96
0.2	20	0.1894	0.0195	0.0689 ± 0.2273	0.30	0.1917	0.0087	0.0096 ± 0.0034	0.80
0.1	20	0.1350	0.0142	0.0509 ± 0.0956	0.14	0.1128	0.0047	0.0067 ± 0.0039	0.56
0	20	0.0669	0.0056	0.0265 ± 0.0255	0.06	0.0231	0.0007	0.0024 ± 0.0015	0.04
0.5	50	0.4949	0.0138	0.0142 ± 0.0136	0.92	0.5003	0.0078	0.0058 ± 0.0006	1.00
0.4	50	0.3708	0.0138	0.0150 ± 0.0149	0.79	0.4028	0.0048	0.0057 ± 0.0005	1.00
0.3	50	0.2967	0.0140	0.0135 ± 0.0087	0.66	0.2957	0.0047	0.0050 ± 0.0007	1.00
0.2	50	0.1833	0.0127	0.0154 ± 0.0145	0.40	0.2031	0.0049	0.0040 ± 0.0009	0.98
0.1	50	0.1011	0.0057	0.0155 ± 0.0061	0.18	0.0978	0.0020	0.0025 ± 0.0008	0.78
0	50	0.0378	0.0021	0.0103 ± 0.0072	0.05	0.0152	0.0002	0.0009 ± 0.0003	0.05

likelihood analysis and $N = 5$, 10, and 20 for the Bayesian analysis. The reason for the different sample sizes was due to limitations of the two methods for either smaller (ML) or larger (Bayesian) sample sizes. In the maximum likelihood framework, samples of size $N = 5$ were occasionally not variable enough in terms of observed topology to obtain unique joint MLEs of γ, t_1, and t_2, with the likelihood surface flat over all values of t_1 and t_2. In the Bayesian framework, the time required to compute the c_j becomes prohibitive when $N > 30$ or so. A total of 100 samples were generated for each of the 18 combinations of γ and sample size under each model. Species tree branch lengths (t_1 and t_2) were set to 1.0 coalescent units to simulate the data.

For the maximum likelihood analysis, the joint MLEs $\hat{\Theta}$ were obtained by searching a grid of 1,000,000 points over the range $0 < \gamma < 1, 0 < t_1 < 5.0$, and $0 < t_2 < 5.0$. The value of 5.0 coalescent units was set as the upper bound on the range of the branch length parameters, because when species tree branch lengths are as long as 5.0 coalescent units, there is a more than a 99% chance that the gene tree matches the species tree, and the model thus predicts no variability in the gene tree topology distribution. Thus the likelihood surface is flat for $t_i > 5.0$. Although occasionally the MLE of t_1 or t_2 was estimated to be on this boundary (this occurred in less than one-third of the cases when $N = 10$, less than 10% of the cases when $N = 20$, and less than 1% of the cases when $N = 50$; this was also less frequent for Model 2 than for Model 1), the estimates of γ were not greatly affected, and these cases are not excluded from the summaries given in Section 3.

In addition, the variance of each MLE was estimated using the observed Fisher information. A likelihood ratio test was also conducted for each sample to test the hypothesis of hybridization ($H_0 : \gamma = 0$ vs. $H_1 : \gamma \neq 0$). The p-value was obtained based on a simulated null distribution of the likelihood ratio statistic. For the Bayesian analysis, both informative and non-informative priors were used. We used the same non-informative priors for all analyses: $\theta = 0.5$, $\alpha_2 = 1$, and $\beta_2 = 1$. The Bernoulli parameter was chosen to be 0.5, so that no particular hypothesis (hybridization v.s. no hybridization) was favored by the prior. $Beta(\alpha_2, \beta_2)$ is a standard uniform distribution which is non-informative in the sense that γ is uniformly distributed between 0 and 1. Parameters in the informative priors were chosen by considering the true value of γ used to simulate the data. Specifically, θ was set to 0.25 if the data were generated with γ being 0.5, 0.4, or 0.3 and to 0.75 if γ was 0.2, 0.1, or 0. The parameters α_2 and β_2 were determined so that the $Beta(\alpha_2, \beta_2)$ distribution was centered at the true γ. Since $Beta(\alpha_1, \beta_1)$ is always concentrated at 0, $\alpha_1 = 0.1$ and $\beta_1 = 1$ were used in all analyses. Each Bayesian run was summarized by computing the posterior mean of γ and the corresponding 95% credible interval.

3. Results

3.1. Maximum likelihood analysis of simulated data

Our model allows us to take two distinct approaches to the detection of hybridization. The first is to test for the presence of hybrid speciation using the likelihood ratio test, and the second is to estimate the proportional contribution to the genome from each parental species. We note that, in either case, the species phylogeny and location of the putative hybridization event are assumed to be known, although speciation times and the timing of the hybridization event need not be specified in advance. To examine the performance of our methods with respect to both of these approaches, we show the results of our simulation study in Table 1, which gives the average MLE of γ, the observed variance of the MLEs, the average and standard deviation of the estimates of the variance of the MLEs, and the estimated power of the likelihood ratio test for each set of 100 simulated samples. In general, the estimates are close to the true values of γ, with more genes leading to higher accuracy. Also, the mean of the estimated variances of $\hat{\gamma}_{MLE}$ over the 100 replicates is generally close to the observed variance, which suggests that estimation using the observed Fisher information is reasonable. The exception to this occurs when the sample size is 10 and Model 1 is considered. In this case, there are a few data sets for which the MLEs of the branch lengths hit the "boundary" of 5.0 coalescent units. When this happens, the variance is estimated to be very large, and the mean is increased. In practice, variance estimation in these cases would not be possible, but for most cases, reasonable estimates are obtained. We discuss possibilities for estimation of γ in these situations where the branch length MLEs occur on the "boundary" further in Section 5. The results of the likelihood ratio test show that the test becomes more powerful as the sample size gets larger. However, the level of the test (which can be measured from the cases where $\gamma = 0$) is larger than the expected 5% in some cases for Model 1, indicating

Table 2
Results of the Bayesian analysis for the simulated data. The first two columns are as in Table 1. Columns 3 and 4 provide information on the prior distributions used in the analysis. To specify an uninformative prior, the prior probability of no hybridization (θ) was set to 0.5, and a uniform distribution over the interval (0,1) was used for γ (labeled as "flat" in column 4). To indicate prior evidence of hybridization, the prior probability of hybridization was set to either 0.75 ($\theta = 0.25$) or 0.25 ($\theta = 0.75$), depending on the true value of γ. In both cases, a Beta distribution centered at the true value of γ (column 1) was used for the $Beta(\alpha_2, \beta_2)$ distribution (labeled as "informative" in column 4). The remaining columns give, for each model, the average of the 100 posterior means for each simulated data set. The posterior probability of no hybridization (θ) is also given. Note that small values of θ correspond to evidence in favor of a role for hybridization.

True γ	Sample size	Prior θ	$Beta(\alpha_2, \beta_2)$	Model 1		Model 2	
				Average posterior mean of γ	Posterior θ	Average posteriormean of γ	Posterior θ
0.5	5	0.25	Informative	0.4498	0.1523	0.4742	0.0938
0.5	5	0.5	Flat	0.4491	0.2956	0.4580	0.2445
0.5	10	0.25	Informative	0.4700	0.1081	0.4784	0.0696
0.5	10	0.5	Flat	0.4320	0.2961	0.4523	0.1980
0.5	20	0.25	Informative	0.4757	0.0777	0.4833	0.0428
0.5	20	0.5	Flat	0.4565	0.2297	0.4714	0.1791
0.4	5	0.25	Informative	0.3762	0.1410	0.3723	0.1207
0.4	5	0.5	Flat	0.4115	0.3278	0.3617	0.2944
0.4	10	0.25	Informative	0.3743	0.1310	0.3807	0.0852
0.4	10	0.5	Flat	0.3580	0.3408	0.3623	0.2430
0.4	20	0.25	Informative	0.3762	0.1043	0.3840	0.0453
0.4	20	0.5	Flat	0.3830	0.2823	0.4031	0.2016
0.3	5	0.25	Informative	0.2826	0.2083	0.3047	0.1461
0.3	5	0.5	Flat	0.3200	0.4004	0.2905	0.3721
0.3	10	0.25	Informative	0.3071	0.1380	0.2912	0.1174
0.3	10	0.5	Flat	0.3193	0.3745	0.3058	0.2877
0.3	20	0.25	Informative	0.2982	0.1507	0.2916	0.0737
0.3	20	0.5	Flat	0.3096	0.3381	0.2862	0.2765
0.2	5	0.75	Informative	0.1997	0.6041	0.1620	0.6103
0.2	5	0.5	Flat	0.2947	0.4216	0.2349	0.4535
0.2	10	0.75	Informative	0.1680	0.6191	0.1901	0.4837
0.2	10	0.5	Flat	0.2209	0.4870	0.2237	0.4078
0.2	20	0.75	Informative	0.1724	0.5603	0.2209	0.3522
0.2	20	0.5	Flat	0.1877	0.5014	0.2034	0.3647
0.1	5	0.75	Informative	0.1457	0.6361	0.1176	0.6326
0.1	5	0.5	Flat	0.2572	0.4496	0.1507	0.5850
0.1	10	0.75	Informative	0.1240	0.6172	0.1183	0.5721
0.1	10	0.5	Flat	0.1870	0.5399	0.1605	0.4899
0.1	20	0.75	Informative	0.1174	0.5985	0.1195	0.4810
0.1	20	0.5	Flat	0.1623	0.5456	0.1801	0.4443
0	5	0.75	Informative	0.0929	0.6822	0.0381	0.6778
0	5	0.5	Flat	0.1734	0.5740	0.0793	0.7108
0	10	0.75	Informative	0.0637	0.6695	0.0359	0.6543
0	10	0.5	Flat	0.1218	0.6338	0.0524	0.7599
0	20	0.75	Informative	0.0432	0.6619	0.0215	0.6697
0	20	0.5	Flat	0.1034	0.6472	0.0977	0.7739

that the test rejects the null hypothesis of no hybridization more often than it should for this model.

It is anticipated that estimation and testing should be more straightforward for Model 2 than Model 1, because a hybridization event between more distantly related lineages should be easier to distinguish from incomplete lineage sorting. This is confirmed by the observation that the MLEs are closer to the true parameters and the variances are smaller for Model 2. In addition, the LRT has higher power for Model 2, and the level of the test is closer to the expected 5% level. We also note that when the sample size is somewhat small (e.g., $N = 10$), the variance of $\hat{\gamma}_{MLE}$ is relatively large, so that the resulting 95% confidence interval for γ might cover a large portion of the interval of [0, 1]. This indicates that the ML method will often be most useful for larger samples.

3.2. Bayesian analysis of simulated data

The results of our Bayesian analyses of the simulated data are summarized in Table 2, which gives the average of the posterior means of γ and the posterior probability of $\epsilon = 1$ for each simulation setting. Fig. 4 is a plot of posterior 95% credible intervals for the set of 100 gene tree samples simulated from Model 2 with $\gamma = 0.5$, which demonstrates that the credible interval performs well with respect to coverage. Overall, the posterior mean and 95% credible interval are reasonable estimates of γ regardless of whether informative or flat priors are used. The posterior probability of $\epsilon = 1$ (no hybridization) is smaller than the prior probability in the cases where the true values of γ are 0.5, 0.4, 0.3, or 0.2 and larger than the prior in the cases of γ truly being 0.1 or 0. As expected, the parameter estimates are better for Model 2 than Model 1. The Bayesian method

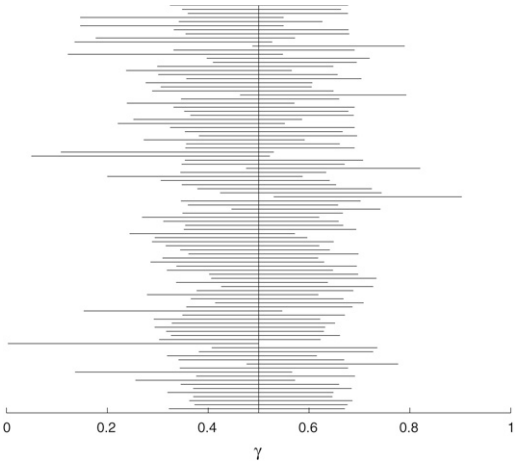

Fig. 4. Posterior 95% credible intervals for γ for 100 simulated data sets when $\gamma = 0.5$ and informative priors are used. In this analysis, 99% of the credible intervals contain the true value of γ.

performs better than the ML method for small samples, especially when informative prior knowledge about γ is available. In this situation, the credible intervals are often much narrower than the corresponding confidence intervals obtained in the likelihood analysis.

4. Empirical data

We next apply the Bayesian method to an empirical data set. In a recent study, Buckley et al. (2006) explored the radiation of New Zealand alpine cicadas, and estimated gene trees separately for four loci. The inferred gene trees show incongruence in the placement of some taxa, especially the species *M. iolanthe*, which is grouped with either *M. campbelli* or with the *M. lindsayi* and *M. myersi* clade. The conflicting position of *M. iolanthe* among the gene trees is possible evidence for a hybrid origin of this taxon. Other evidence suggests the possibility of a hybrid origin as well. For example, all of the four species involved tend to inhabit low-altitude habitats, have morphologically similar genitalia, are of a similar body size, and have similar courtship songs (Buckley et al., 2006). However, Buckley et al. (2006) note that hybridization is not a common event in *Maoricicada*, and that no shared alleles between pairs of sympatric *Maoricicada* species have been observed.

To determine whether the incongruence in the placement of *M. iolanthe* is possibly due to hybridization, we consider a model where *M. iolanthe* is a hybrid of the parental species *M. campbelli* and *M. lindsayi*, with *R. leptomera* as an outgroup species. We consider two possible methods for handling branch lengths in the hybrid species tree. First, we choose branch lengths based on previous estimates of divergence dates of New Zealand cicadas (Arensburger et al., 2004; Buckley et al., 2006) and fix these prior to a Bayesian analysis. We also consider obtaining MLEs for these species tree branch lengths, and fixing branch lengths at the MLEs. In each case, two sets of priors were used in the Bayesian analysis. The first set is heavily weighted towards hybridization by setting $\theta = 0.1$, $\alpha_1 = 0.1$, $\beta_1 = 1$, $\alpha_2 = 10$, and $\beta_2 = 10$. The second set is less-informative with $\theta = 0.1$, $\alpha_1 = 0.1$, $\beta_1 = 1$, $\alpha_2 = 1$, and $\beta_2 = 1$.

In order to monitor convergence, we ran multiple chains with the same prior from different starting points. The MCMC appeared to be fast to converge and results obtained from multiple runs are similar to each other, regardless of the starting point. After 100,000 iterations, the PSRFs were <1.006 in all cases. Each run took approximately 15 min on a Sun Blade 2500 (utilizing one of two available 1.6 Hz UltraSPARC IIIi processors under the Solaris operating system).

The trace plots and posterior histograms of γ with different priors and different methods for handling the species tree branch lengths are shown in Fig. 5. Using the informative priors and independently specified species tree branch lengths, the posterior mean of γ is 0.4927, with 95% credible interval (0.2541, 0.7019), and the posterior probability of no hybridization is 0.0302, much smaller than the prior probability of 0.1. The posterior distribution of γ is highly concentrated at 0.5 (see Fig. 5(b)). Because the posterior probability of no hybridization has decreased substantially in comparison with the prior probability, and because the 95% credible interval for γ does not include either 0 or 1, these results indicate a substantial role for hybridization in the formation of the species *M. iolanthe*. Using the non-informative priors and independently specified species tree branch lengths, the posterior mean of γ is 0.4906. The posterior probability of $\epsilon = 1$ is 0.0398 which is also less than the prior value. Although the 95% credible interval for γ is (0.0379, 0.9260), which is very spread out, the posterior distribution with a single peak at 0.5 (Fig. 5(d)) shows some evidence for hybridization.

Similar results are obtained when species tree branch lengths are fixed at their MLEs. In the case of informative priors, the posterior mean of γ is 0.4975, with 95% credible interval (0.2933, 0.6952), and the posterior probability of no hybridization is 0.0156. When non-informative priors are used, the posterior mean of γ is 0.5034, with 95% credible interval (0.1427, 0.8561), and the posterior probability of no hybridization is 0.0233. Thus we see that both methods of handling species tree branch lengths lead to very similar inferences concerning γ.

5. Discussion

The ML and Bayesian methods produce similar estimates for the simulated data in our study. However, the ML method can perform poorly for small samples. For example, when the sample size is small ($\leq$10), the 95% confidence interval of γ might cover much of the interval (0, 1), while the Bayesian method with informative priors performs better in such cases. From the empirical data, we see that the Bayesian estimates are reasonable, even for only four genes with non-informative prior distributions. In addition, the Bayesian approach provides a framework for incorporating prior information about the parameters. The advantage of using priors is particularly valuable when the gene tree data are limited but a wealth of information is available about hybridization. For example, hybridization may be observed in nature or might be demonstrated in the lab, and in such cases it is worthwhile to consider these prior beliefs about hybridization in the analysis through use of a prior that puts weight on values of γ that indicate hybridization. However, due to the calculation of the c_j's in Eq. (11), the computing time of the Bayesian algorithm increases exponentially as the sample size gets larger. Fortunately, for large samples, the ML method is shown to be a reliable estimation procedure.

The implications of the simulation results in this article generalize directly to trees with more taxa, because our model can be easily extended to trees of any size. Regardless of the number of taxa, a hypothesized hybridization event gives rise to two sets of species relationships, one with probability γ and the other with probability $1 - \gamma$. The coalescent process then operates under each relationship, and the likelihood is easily computed, since the gene tree distribution needed is available for any number of taxa (Degnan and Salter, 2005). One possible complication is that the number of branch lengths in the species phylogeny grows as the number of taxa increases, and thus the grid search implemented here for the case of four taxa will become inefficient. However, it is likely that algorithms to optimize branch lengths could be easily implemented using tools analogous to those used to compute maximum likelihood branch lengths along single gene phylogenies, though we have not explored this possibility specifically. Additionally, it is possible to set all or a subset of the branch lengths to fixed values prior to performing the analysis, if independent information concerning these values is available. We have observed that estimates of the γ parameter are relatively robust to the particular values of the branch lengths used in our four-taxon examples.

It is also straightforward to consider more than one hybridization event by allowing a separate γ parameter for each putative occurrence of hybrid speciation. Under some restrictions on the set of hybridization events that are allowed to occur along the tree, it is possible to prove asymptotic MLE results analogous to those presented here for a single hybridization event. For example, when each branch on the species phylogeny is only allowed to "participate" in a single hybridization event, we have derived results analogous to those in Eqs. (3)–(6). Either the maximum likelihood method or the Bayesian method can then be extended

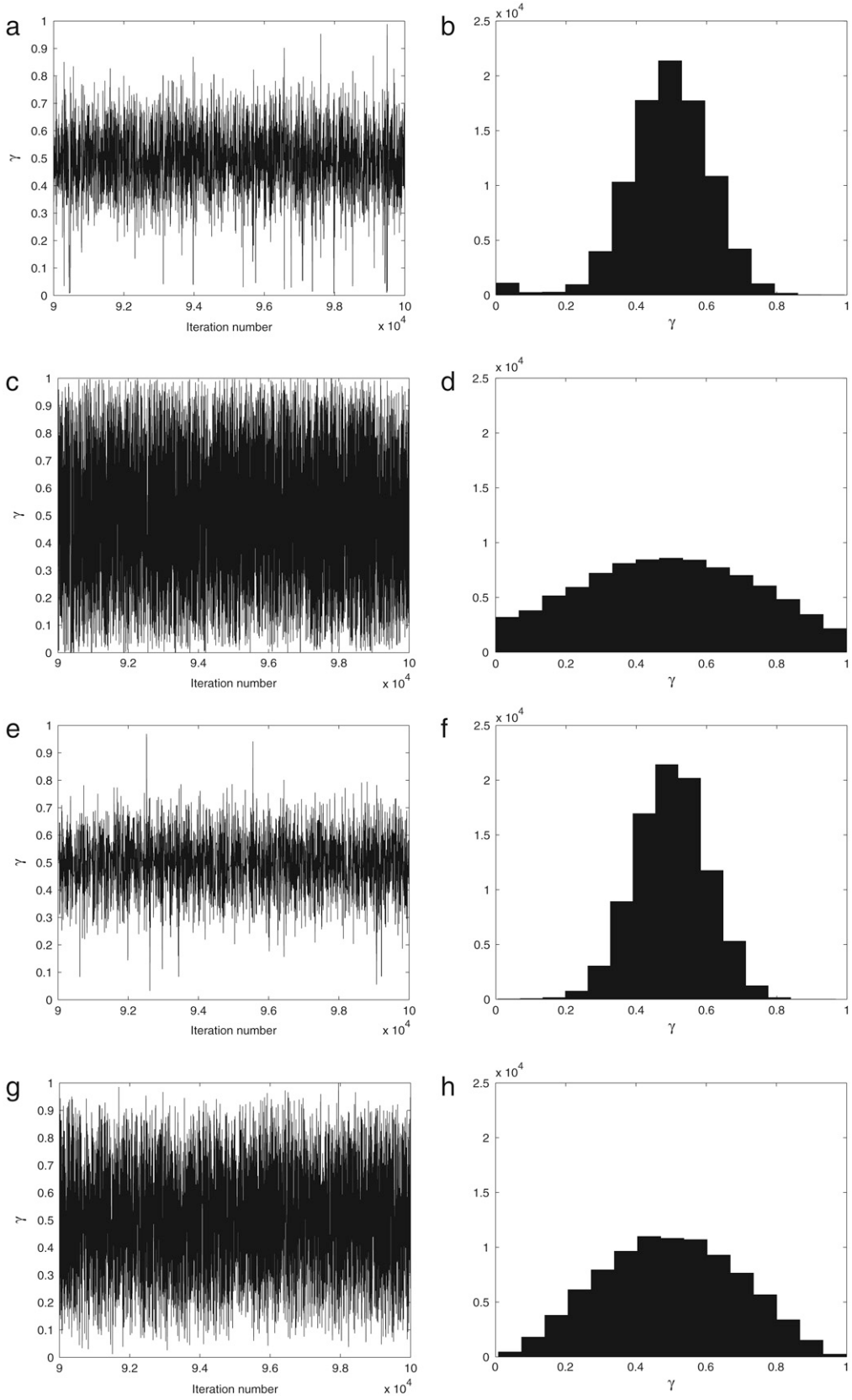

Fig. 5. Results of the analysis of the real data of Buckley et al. (2006) using informative (a, b, e, f) and uninformative (c, d, g, h) prior distributions (see text for details). (a), (c) Trace plots for γ using informative and uninformative prior distributions, respectively, when branch lengths are fixed values independently specified (only the last 10,000 iterations of the chain are plotted to simplify the visual display; trace plots of all iterations do not show any deviation from the above patterns); (b), (d) Histograms of sampled values of γ using informative and uninformative prior distributions, respectively, when branch lengths are fixed at values independently specified (all iterations are included); (e), (g) Trace plots for γ using informative and uninformative prior distributions, respectively, when branch lengths are fixed at the MLEs (only the last 10,000 iterations of the chain are plotted to simplify the visual display; trace plots of all iterations do not show any deviation from the above patterns); (f), (h) Histograms of sampled values of γ using informative and uninformative prior distributions, respectively, when branch lengths are fixed at the MLEs (all iterations are included).

to carry out estimation or hypothesis testing. We have not yet explored this model in more complicated settings, for example, cases in which taxa formed by hybridization subsequently participate in other hybridization events, though we expect such situations to be much more difficult to detect using only information in gene tree topologies.

Here we have considered the problem of estimation of hybrid speciation in species phylogeny based on incongruent gene topologies. Our model assumes that the location of hybridization events in a given species tree is known, and also that gene tree topologies are known without error. A more general method would incorporate variability in each of these quantities as well. Another limitation of the current study is that the branch lengths of the gene trees are not used in the analysis. We observe that the method performs remarkably well, in spite of the fact that this additional branch length information is not incorporated into the likelihood function. Our future research involves the development of a Bayesian method to allow for simultaneous estimation of all of these quantities, which will lead to more accurate inference of phylogenies subject to both hybridization and incomplete lineage sorting due to deep coalescence. In spite of the fact that only a subset of the available data (only gene tree topologies) is used in the analysis, the method proposed here demonstrates good ability with respect to both testing and estimation of hybridization leading to speciation. It is therefore an important step in the development of models that simultaneously incorporate both hybridization and incomplete lineage sorting into phylogenetic analysis in a likelihood-based framework.

Acknowledgments

We thank Radu Herbei for making us aware of the work of George and McCulloch and for helpful discussions concerning our MCMC implementation. We thank Lisle Gibbs and two anonymous reviewers for useful comments on an earlier version of this manuscript, and James Degnan for helpful discussions concerning modeling hybridization. This work was partially supported by National Science Foundation grant DMS 0702277 to L.S.K.

Appendix. Computation of c_j's

The probability of an i.i.d sample of gene trees given the hybridization model, i.e., the likelihood, is

$$L(\gamma|Data) = \prod_{i=1}^{N} P(g_i|\gamma) = \prod_{i=1}^{N} \{\gamma P(g_i|\tau_1) + (1-\gamma)P(g_i|\tau_2)\}.$$

Let $a_i = P(g_i|\tau_1) - P(g_i|\tau_2)$ and $b_i = P(g_i|\tau_2)$. The likelihood is then written as

$$\begin{aligned} P(Data|\gamma, \epsilon) &= \prod_{i=1}^{N} (a_i\gamma + b_i) \\ &= c_N\gamma^N + c_{N-1}\gamma^{N-1} + \cdots + c_1\gamma + c_0 \\ &= \sum_{j=0}^{N} c_j\gamma^j. \end{aligned}$$

Here,

$$c_N = \prod_{i=1}^{N} a_i$$

$$c_{N-1} = \sum_{k=1}^{N} \left(\frac{b_k \prod_{i=1}^{N} a_i}{a_k} \right)$$

$$c_{N-2} = \sum_{k=1}^{N} \sum_{l=k+1}^{N} \left(\frac{b_k b_l \prod_{i=1}^{N} a_i}{a_k a_l} \right)$$

.

.

.

$$c_0 = \prod_{i=1}^{N} b_i.$$

References

Arensburger, P., Buckley, T.R., Simon, C., Moulds, M., Holsinger, K.E., 2004. Biogeography and phylogeny of the New Zealand cicada genera (Hemiptera: Cicadidae) based on nuclear and mitochondrial DNA data. J. Biogeography 31, 557–569.

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

Baack, E.J., Rieseberg, L.H., 2007. A genomic view of introgression and hybrid speciation. Curr. Opin. Genetics Dev. 17, 1–6.

Buckley, T.R., Cordeiro, M., Marshall, D.C, Simon, C., 2006. Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (*Maoricicada dugdale*). Syst. Biol. 55 (3), 411–425.

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

Dowling, T.E., Secor, C.L., 1997. The role of hybridization and introgression in the diversification of animals. Annu. Rev. Ecol. Syst. 28, 593–619.

Edwards, S.V., Liu, L., Pearl, D.K., 2007. High resolution species trees without concatenation. PNAS 104, 5936–5941.

Gauthier, O., Lapointe, F.J., 2007. Hybrid and phylogenetics revisited: A statistical test of hybridization using quartets. Syst. Botany 32 (1), 8–15.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7, 457–511.

George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. J. Am. Stat. Assoc. 88 (423), 881–889.

Guigo, R., Muchnik, I., Smith, T.F., 1996. Reconstruction of ancient molecular phylogeny. Mol. Phylogenet. Evol. 6, 189–213.

Hennig, W., 1966. Phylogenetic Systematics. Illinios Press, Urbana-Champaign.

Holder, M.T., Anderson, J.A., Holloway, A.K., 2001. Difficulties in Detecting Hybridization. Syst. Biol. 50 (6), 978–982.

Hudson, R.R., 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution 37, 203–217.

Kingman, J.F.C., 1982. The coalescent. Stoch. Proc. Appl. 13, 235–248.

Legendre, P., Makarenkov, V., 2002. Reconstruction of biogeographic and evolutionary networks using reticulograms. Syst. Biol. 51 (2), 199–216.

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.

Maddison, W.P., 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. TREE 20 (5), 229–237.

Mallet, J., 2007. Hyrbid speciation. Nature 446, 279–283.

Medigue, C., Rouxel, T., Vigier, P., Henaut, A., Danchin, A., 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. J. Mol. Biol. 222, 851–856.

Mossel, E., Roch, S., 2008. Incomplete lineage sorting: Consistent phylogeny estimation from multiple loci. http://arxiv.org/abs/0710.0262.

Nei, M., 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Nordborg, M., 2001. Coalescent theory. In: Balding, D., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley, Chichester, pp. 179–212.

Pamilo, P., Nei, M., 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5 (5), 568–583.

Posada, D., 2002. Evaluation of methods for detecting recombination from DNA sequences: Empirical data. Mol. Biol. Evol. 19, 708–717.

Rannala, B., Yang, Z., 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164, 1645–1656.

Rieseberg, L.H., 1997. Hybrid origins of plant species. Annu. Rev. Ecol. Syst. 28, 359–389.

Rieseberg, L.H., Morefield, J.D., 1995. Character expression, phylogenetic reconstruction, and the detection of reticulate evolution. In: Hoch, P.C., Stephenson, A.G. (Eds.), Experimental and Molecular Approaches to Plant Biosystematics. Missouri Botanical Garden, St. Louis, pp. 333–353.

Rieseberg, L.H., Wendel, J.F., 1993. Introgression and its consequences in plants. In: Harrison, R.G. (Ed.), Hybrid Zones and the Evolutionary Process. Oxford University Press, pp. 70–109.

Rosenberg, N.A., 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61, 225–247.

Sang, T., Zhong, Y., 2000. Testing hybridization hypotheses based on incongruent gene trees. Syst. Biol. 49 (3), 422–434.

Self, S.G., Liang, K.Y., 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. 82 (398), 605–610.
Strimmer, K., Moulton, V., 2000. Likelihood analysis of phylogenetic networks using directed graphical models. Mol. Biol. Evol. 17 (6), 875–881.
Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105, 437–460.
Takahata, N., 1989. Gene genealogy in three related populations: Consistency probability between gene and population trees. Genetics 122, 957–966.
Takahata, N., Nei, M., 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110, 325–344.
Tateno, Y., Nei, M., Tajima, F., 1982. Accuracy of estimated phylogenetic trees from molecular data. I. distantly related species. J. Mol. Evol. 18, 387–404.
Tavaré, S., 1984. Line-of-descent and genealogical processes, and their applications in population genetics models.. Theor. Popul. Biol. 26, 119–164.
Than, C., Ruths, D., Innan, H., Nakhleh, L., 2007. Confounding factors in HGT detection: Statistical error, coalescent effects, and multiple solutions. J. Comp. Biol. 14 (4), 517–535.
Valdez, A.M., Pinero, D., 1992. Phylogenetic estimation of plasmid exchange in bacteria. Evolution 46 (3), 641–656.
Xu, S., 2000. Phylogenetic analysis under reticulate evolution. Mol. Biol. Evol. 17 (6), 897–907.