Bayesian phylogenetic analysis for diploid and allotetraploid species networks

Graham Jones

May 6, 2013

Abstract

Allopolyploid species are formed by genome doubling after hybridization between otherwise intersterile parental species. Allopolyploidy is the second most common speciation mechanism in land plants, after ordinary speciation. Here we describe and evaluate a Bayesian approach to the phylogenetic analysis of species relationships when both ordinary speciation and allopolyploidy are present. The approach takes incomplete lineage sorting into account using the multi-species coalescent model, and extends this to deal with the extra complications due to allopolyploidy. The number of hybridizations is not assumed, which means that the number of parameters varies and a reversible-jump MCMC algorithm is needed to sample from the posterior. The main restriction is that only diploids and allotetraploids are considered. The model is implemented in the BEAST framework. Simulations show that the topology of the network can be reliably inferred along with estimates of other parameters. The models are demonstrated on previously analyzed data the genus *Silene* (Caryophyllaceae).

1 Introduction

Speciation may involve hybridization or genome doubling. Ordinary speciation, which involves neither, is the most common and the most studied. When it occurs repeatedly it forms a branching process, and the result is a binary tree which grows as branches at the tips split into two. Speciation which involves only genome doubling is called autopolyploidy, and it can still be seen as part of a branching process. Speciation which involves hybridization, but no change in genome size is called homoploid hybridization or recombinational speciation; the genome of the new species is a 'mixture' or 'mosaic' of the genomes of the two parental species. Finally, in allopolyploidy, individuals belonging to two different species produce one or more hybrid individuals which then undergo genome doubling and form a new species. The hybrid individuals would (at least in most cases) be infertile without this genome doubling, because the chromosomes of the two parental species are too different to recombine during meiosis. In this case the genome of the new species is the 'sum' of the genomes of the two parental species.

In general, homoploid hybridization, autopolyploidy, and allopolyploidy are rare in comparison to ordinary speciation. Homoploid hybridization is hard to detect but has been found in a few cases. On the other hand polyploids (species with doubled or other multiples of genomes compared to close relatives) are common in plants, and also occur in animals and fungi. Around half of flowers and 95% of ferns are polyploids. Polyploids are relatively easy to detect, but it is harder to determine whether their origin was autopolyploidy or allopolyploidy, or to determine how much ordinary speciation has occurred since their origin. Autopolyploidy is much more common than allopolyploidy. Within plants at least, allopolyploidy is the most important mechanism for generating new species apart from ordinary speciation.

The purpose of this paper is to describe and evaluate a Bayesian approach to the phylogenetic analysis of species relationships when both ordinary speciation and allopolyploidy are present. It builds on the work in [10] which was restricted to a single hybridization. The main restriction here is that only diploids and allotetraploids are considered. Thus we assume that the species being analyzed have un-

dergone at most one round of allopolyploidization since the root of the network. We also assume that within the allotetraploids, there is no recombination between sequences from different parental species. This means that all the sequences can be seen as the result of the evolution of diploid genomes, but after hybridization, node times and population sizes are shared.

The evolutionary history can be represented as a network or as a multi-labeled tree (MUL-tree), which is a binary tree in which more than one tip may be labeled by the same species. Both these representations omit information about extinct species; they are reconstructions from extant taxa. An example is shown in Figure 1. This shows a scenario which results in two extant diploid species a and b, and three extant allotetraploid species x, y, and z. Reading these diagrams from left to right, there are three ordinary speciations resulting in four diploid species before the first hybridization. Then two of these hybridize to form allotetraploid species x which continues to present. This is followed by two more ordinary speciations of diploid species, and a second hybridization to produce an allotetraploid species, which then undergoes ordinary speciation to produce y and z. The diploid species which contribute to the hybridizations become extinct before present time, leaving a and b. Note that it is straightforward to convert the network representation into a unique multi-labeled tree representation. There are algorithms for the reverse operation [7] but the network obtained is not in general unique.

There are several problems to deal with in the phylogenetic analysis. In common with the inference of species trees in which there is only ordinary speciation, the issue of incomplete lineage sorting cannot be ignored. Thus we need to simultaneously estimate the gene trees and the species network into which they fit. We use a generalization of the multispecies coalescent model to deal with this. Secondly, when the DNA from allotetraploid organisms is sequenced, it is not possible a priori to assign sequences to their parental diploid species. Thus there is an ambiguity in the labeling of the sequences which is not normally present. These two issues were dealt with in [10], but there a single hybridization was assumed. Here we infer the number of hybridizations. This means that inference must explore a space of species networks in which the number of parameters (node times and population size parameters) varies. We use a reversible-jump MCMC process to explore this space.

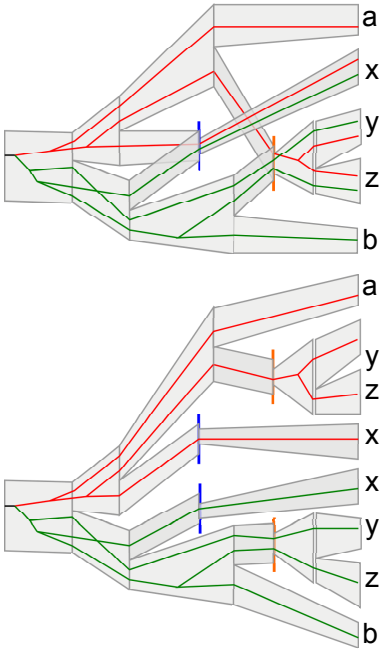

Figure 1: Top: a species network for two diploid species a and b, and three allotetraploid species x, y, and z. The widths of the gray 'tubes' indicate population sizes. The network contains a gene tree, with red and green branches indicating different parental species. The blue line indicates one hybridization, the orange lines another. Bottom: the same scenario represented as a multi-labeled tree.

There has been a considerable amount of work on hybridization, often focussed on homoploid hybridization [1–3, 12]. However these do not take advantage of the particular patterns that occur in allopolyploidy. Much previous work on allopolyploidy [7, 8, 13–15] has used *ad hoc* instead of a statistical approaches. There is more discussion of related work in [10].

2 Model and priors

There is a large amount of notation which we collect here. Let the number of diploid species be d and the number of allotetraploid species be m. The network is denoted by W and the multi-labeled species tree derived from it is M_W. For a given network state,

let h be the number of hybridizations. Suppose the ith allotetraploid subtree has m_i tips ($1 \leq i \leq h$). Then $m = \sum_{i=1}^{h} m_i$. The population size parameters are denoted by the vector θ. The parameter η is a scaling factor for the population sizes, appearing in a hyperprior for θ. The number of gene trees is denoted by G. The topology and set of node times for the ith gene tree is denoted by τ_i ($1 \leq i \leq G$). All the other parameters belonging to the ith gene tree are denoted by α_i; these are parameters for site rate heterogeneity, substitution model, branch rate model, and root model. Thus (τ_i, α_i) gives all the parameters for the ith gene tree. The permutations of sequences within polyploid individuals for the ith gene is denoted by γ_i. This parameter is the main addition to the usual formula for the multispecies coalescent. We only deal with tetraploids here, so γ_i consists of transpositions ('flips') of two sequences. The sequence data for the ith gene is denoted by y_i. We set $\tau = (\tau_1, ...\tau_G)$, and similarly for α, γ, y.

2.1 Model

The formula for the posterior density for the AlloppNET model is similar to that used in [10] and is given by

$$\begin{aligned} f(W, \theta, \tau, \alpha, \gamma | y) \quad \propto \quad & f_W(W|\lambda) f_\lambda(\lambda) \times \\ & f_\theta(\theta|\eta) f_\eta(\eta) \times \\ & f_\gamma(\gamma) \times \\ & \prod_{i=1}^{G} f_\tau(\tau_i | M_W, \theta, \gamma_i) \times \\ & \prod_{i=1}^{G} \Pr(y_i | \tau_i, \alpha_i). \end{aligned} \tag{1}$$

The network prior is $f_W(W|\lambda) f_\lambda(\lambda)$. The choice of this prior poses some problems and is described in a separate section.

The population size prior $f_\theta(\theta|\eta) f_\eta(\eta)$ is for the vector of population size parameters θ. There is a mapping from θ to values at nodes in the network, and to just after each hybridization event. There are $3d + 4m + h - 2$ parameters; note that h varies. Details of the mapping are in the SI. **TODO**. The population sizes are assumed to vary linearly along edges in the network, except that a instantaneous change is allowed at hybridization events. In the analyses done in this paper, the priors for θ used were similar to those typically used by *BEAST. An independent gamma distribution is assumed for each component of θ. The shape parameter is 4 for the populations at the tips, 1 for just after hybridizations, and 2 for the rest. The scale parameter for all these gamma distributions is the hyperparameter η. The hyperprior f_η for η is described later.

The permutation prior $f_\gamma(\gamma)$ is a discrete distribution on the set of sequence assignments. This is assumed to be uniform here, and thus could be omitted without affecting the inference.

The term $f_\tau(\tau_i | M_W, \theta, \gamma_i)$ provides the probability of τ_i, when permuted by γ_i, fitting into the multi-labeled species tree M_W with population sizes determined by θ. The value of γ_i determines how the sequences for the ith gene are assigned to tips in M_W. Note that this probability does not depend on α_i. Apart from this extra complexity due to the permutations, the value of $f_\tau(\tau_i | M_W, \theta, \gamma_i)$ is given by the multispecies coalescent, as used in [18], [6] and elsewhere.

The term $\Pr(y_i | \tau_i, \alpha_i)$ is the probability of the data for the ith gene given the ith gene tree and other parameters α_i. Regarded as a likelihood, it is the usual 'Felsenstein likelihood'. Here α_i contains the substitution model parameters, branch rate model parameters, and site rate heterogeneity model parameters for the ith gene tree. In this paper, we used the HKY substitution model, and assumed strict clock branch rates, and no site rate heterogeneity. The clock rate for one gene was fixed to 1.0, and the others were estimated.

The priors for the population parameter η and the parameter λ appearing in the network prior, and the priors for relative clock rates were diffuse lognormals.

2.2 Network Prior

There are two difficulties. Firstly, there is very little empirical evidence to guide the choice of prior. Secondly, there is little in the way of theory about probability densities on networks, especially when the number of nodes can vary as it does here. The situation can be contrasted with that of species delimitation (eg [20]) where species trees with different numbers of nodes are considered. In that case, the theory of birth-death process provides normalized densities for species trees of different sizes which can then be used to determine the Hasting ratios in the

MCMC algorithm. But for allopolyploid networks no such densities are known. We therefore resort to writing down a formula for an unnormalized density for each possible size of network, and estimating the properties of the prior by sampling from it.

There are $\sum_{i=1}^{h}(m_i-1) = m-h$ internal nodes in the allotetraploid trees. There are h hybridization times, and the diploid history has $d+2h-1$ internal nodes. The total number of parameters (node times and hybridization times) which are operated on by the reversible jumps is thus $n := d + m + 2h - 1$. If the ratios between different models (different h) is to be the same regardless of λ, then the density must reflect this. The formula we use is

$$\left(7\sqrt{d}\right)^{-h} \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} t_i\right) \qquad (2)$$

where the t_i are all the node times and hybridization times. The factor $(7\sqrt{d})^{-h}$ was chosen experimentally so that the marginal distribution over $h \in \{1, \ldots ...m\}$ was approximately uniform. Note that λ something like a diversification rate.

3 MCMC implementation

The network can be represented as a set of allotetraploid subtrees and a 'diploid history', as in Figure 2. The diploid history is an ordinary tree with some tips ('hybridization tips') having nonzero times. There is one pair of hybridization tips for each hybridization and both tips have the same time, which is the time at which hybridization occurred and hence also the time of origin for an allotetraploid subtree. We will also refer to the external branches in the diploid history which lead to the hybridization tips as the 'legs' of the corresponding allotetraploid subtree.

In this section we describe the MCMC operators (moves) used, and the choice of initial state. There are five novel types of move which are particular to allopolyploid networks:

1. Change a hybridization time.
2. Change an allotetraploid subtree, tipwards of the hybridization
3. Change the diploid history, rootwards of hybridizations.
4. Change the number of allotetraploid subtrees, that is, the number of hybridizations.

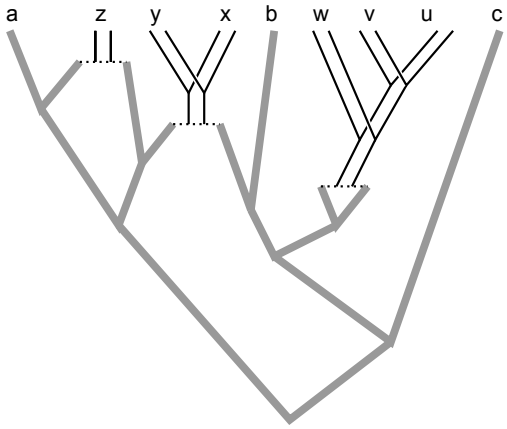

Figure 2: Network with three allotetraploid subtrees (pairs of thin black lines) and the diploid history (thick gray lines). The allotetraploid subtrees have sets {z}, {x,y}, and {w,v,u} of extant species. The diploid history has extant species a,b, and c, plus three pairs of tips which each lead to an allotetraploid subtree.

5. Change the assignment of sequences within polyploid individuals.

The first is straightforward. The next two have much in common and are described next.

3.1 Allotetraploid subtrees and diploid history

Like *BEAST, we use a MCMC move for the species tree based on the ideas of [16]. One reason for using this move is that it can be constrained to keep the species tree compatible with the gene trees. This MCMC move randomly assigns 'left' and 'right' labels to the immediate descendants of each node, to produce an oriented tree [4] and then alters a node height. In our situation this must take into account the assignment of sequences within allotetraploid individuals. Otherwise, within a single allotetraploid subtree the situation is very similar to that in *BEAST.

The same type of move can be adapted to deal with the diploid history. There are three types of constraint on the new height. Firstly, there are constraints from the gene trees as in the tetraploid subtrees, but the calculation is more complex. In order to calculate the sets of sequences belonging to the left and right subtrees (in the MUL-tree) of a par-

ticular node in the diploid history, it is necessary to visit the tetraploid subtrees which are attached to the hybridization tips of the diploid history. Secondly, there are lower bounds on the new height due to the fact that the hybridization tips have nonzero height. In the oriented tree, this amounts to ensuring that the new height does not become smaller than either of the heights of adjacent nodes. (Nodes adjacent to internal nodes are always tips in the left-right ordering.) Thirdly there are constraints to keep the root as a diploid. If node to change height is the root, and the second highest node is to left or right of all diploids, then the root must stay the root: there is a lower limit which is the height of second highest node. If the node to slide is not the root, and is to left or right of all diploids, then it must not become the root: there is an upper limit which is the root height.

3.2 The number of hybridizations

This is the most complex move. Sampling all values of h can be done by repeatedly changing h to $h-1$ or $h+1$, and that can be done by splitting one tetraploid subtree into two and merging two into one. The difficult part is making the moves reversible, so that the probability of a move going from one network state A to another B is balanced by a reverse move. When the MCMC move for changing h is chosen, a split or a merge is chosen with equal probability. If a split is chosen, but no splits are possible, no move is made; the same network state is sampled again. Likewise, if a merge is chosen, but no merges are possible, the same network state is sampled again.

Splitting (going left to right in the top section of Fig 3). Any tetraploid subtree with more than one tip can be split. One, T, is chosen at random. The two child nodes of the root of T become the roots of the two new tetraploid subtrees. The child nodes are not treated symmetrically in the move, so both orderings of the child nodes is treated as candidates. There are thus twice as many candidate splits as tetraploid subtrees. The steps are:

1. Split T into T1 and T2 and create a new hybridization height for T1 between the root height of T1 and the root height of T.
2. Create a new hybridization height for T2 between the root height of T2 and the root height of T. Create two ancestor nodes for the hybridization tips, one re-using the root height of T and the other between this time and the minimum of the limits imposed by the gene trees and the height of the node that will become its ancestor node.
3. Join up the topology in the diploid history. Note that they could join the diploid history in many ways but a particular way is always chosen, so that the two new subtrees 'share legs' as shown. Note also that it is necessary to keep track of which of the two 'copies' of a tetraploid subtree is which (ie, the left leg of one must correspond to the left leg of the other).

The most difficult parameter is the new node height when splitting, which is below the root height of the subtree. This can conflict with gene trees, so the sets of (species,sequence) pairs have to be found for the two child nodes of the new node, and the gene trees examined to find the most recent coalescence that conflicts with the new node.

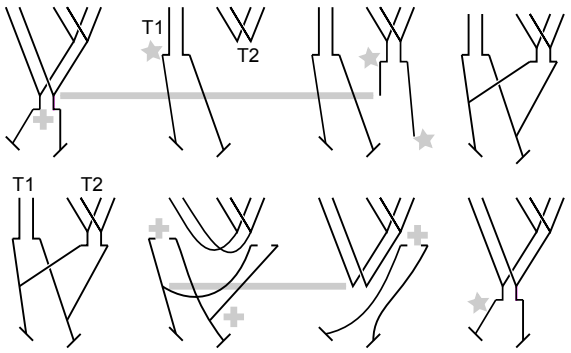

Figure 3: MCMC moves to change the number of tetraploid subtrees. Top: splitting. Bottom: merging. The two gray horizontal lines show times that are reused. Gray crosses indicate times that are about to disappear. Gray stars show times which have just been created.

Merging (going right to left in the bottom section of Fig 3) must be the reverse of splitting. So two tetraploid subtrees can only merge if they have a configuration like that in the figure, ie 'sharing legs'. It is not necessary that the two nodes at the bottom of the figure be different. A list of possible pairs of tetraploid subtrees is made, and if there any suitable pairs, one pair T1,T2, is chosen at random, and the merge is carried out. The steps are:

1. Merge T1 and T2 into T. The root height of T becomes the height of the most recent ancestor to a hybridization tip.
2. Remove the hybridization tips for T1 from the diploid history. This loses a hybridization height and a node height. The latter requires finding the limit from gene trees for Hastings ratio.
3. Give the hybridization tips for T1 new heights below the root height of T, and join up.

Note that the limit from gene trees must be calculated for the lost node height (as when splitting), in order to calculate the Hastings ratio.

We use the theory developed in [5] to calculate the Hastings ratios for splitting and merging moves. Consider a splitting move; the merging case is similar. Then equation (7) of [5] provides the acceptance ratio

$$\min\left\{1,\ \frac{p(S,\theta^{(S)}|y)j(S,\theta^{(S)})q_S(u^{(S)})}{p(M,\theta^{(M)}|y)j(M,\theta^{(M)})q_M(u^{(M)})}\,J\right\} \quad (3)$$

where

$$J = \left|\frac{\partial(\theta^{(S)},u^{(S)})}{\partial(\theta^{(M)},u^{(M)})}\right|$$

is the Jacobian and S (for 'split') and M (for 'merged') replace Green's 1 and 2. Here y is the data, and $p(S,\theta^{(S)}|y)$ and $p(M,\theta^{(M)}|y)$ the usual Bayesian posteriors. The term $j(S,\theta^{(S)})$ gives the probability of choosing the splitting move and $j(M,\theta^{(M)})$ of choosing the reverse merging move. The term $\theta^{(S)}$ is a vector of node and hybridization times for the network with the split case, and $\theta^{(M)}$ is a vector of node and hybridization times for the network with the merged case. The vectors $u^{(S)}$ of length 3 and $u^{(M)}$ of length 1 provide the extra parameters created when doing the jump. The function q_S is the density of the distribution from which $u^{(S)}$ is sampled; likewise q_M.

In our case $u^{(S)}$ and $u^{(M)}$ can be generated by independent sampling from the uniform distribution on $[0,1]$ for each dimension. In this case $q_S(u^{(S)})$ and $q_M(u^{(M)})$ are both 1 and can be omitted from the formula. The new parameters are then derived from these values, as functions of the other old parameters. In the present case, all these function are linear functions mapping $[0,1]$ to a suitable range, the range being some function of the other parameters $\theta^{(S)}$ or $\theta^{(M)}$. The probabilities of the moves being chosen, namely $j(\mathrm{M},\theta^{(M)})$ and $j(\mathrm{S},\theta^{(S)})$ must also be taken into account.

When the number of hybridizations changes, the number of population parameters also changes. It increases by one for each hybridization. So in a split a new population parameter is added, and in a merge one is removed. When one is added, it is sampled from the prior for the population. The contribution to the Hastings ratio is calculated from the value of the density of the population prior at the new parameter value when splitting, or at the lost value when merging.

3.3 Assignment of sequences within allotetraploid individuals

A uniform prior on the possible assignments is used, so the Hastings ratios are all 1. The reassignment moves need to visit all possible 'flips' for each gene in each tetraploid individual. This is easy to arrange in a mathematical sense. The only difficulty is choosing combinations of flips which have good mixing properties. As usual with MCMC algorithms, this requires experimentation.

Three types of move have been implemented. The first flips the assignment of a single gene in a single individual. The second works within a single gene tree, and chooses a random node within it, and roughly speaking, flips a clade of individuals. The details are in the code.

The third type of move was introduced to avoid the MCMC chain getting stuck in a particular situation. The merging move which reduces the number of hybridizations by one, requires the left legs to be 'shared', and the right legs to be shared; left leg shared with right leg will not do. However the network can get into a situation where the legs are 'reversed' *and* the sequence assignments of the genes at the tips of the two candidate are opposite to one another. This makes a state which may be very difficult to leave. The third move flips sequence assignments of all genes of all individuals of all species in a tetraploid subtree, *and* switches the legs around, swapping left and right. This takes the network to an equivalent state with the same likelihood, but one from which the merging move can operate.

3.4 Initial state

A random initial state for the species network is chosen as follows.

1. The tetraploid species are partitioned into one or more groups using the Chinese restaurant process [11, 17].
2. Trees from a Yule process are generated for each of the groups of tetraploid species.
3. The diploid history is constructed in a manner similar to the Yule process working backwards in time, with some modifications. Each diploid species is a tip and there are two hybridization tips for each tetraploid subtree. Two subtrees are repeatedly chosen from those available and joined into a subtree. When two nodes are selected for joining, the choice is constrained so that the diploids do not all merge while there are still tetraploids left to merge. Also, the height of the root of the new subtree has to be made earlier than either of the nodes chosen for joining (which would not happen automatically, since the hybridization tips have nonzero height).

4 Simulations

All genes have length 500 (except in the case of no data). Population sizes are 100,000 individuals (hence 200,000 gene copies per diploid genome) at the tips, and at rootward ends of branches, and 200,000 individuals at tipward ends of internal branches and at the root. Strict clock branch rates, no site rate heterogeneity, and equal clock rates for all genes were assumed. The HKY substitution model was assumed and parameter kappa was set to 3, and the frequencies set to .3 for A and T, and .2 for C and G (Seq-Gen was called with parameters `-t3.0 -f0.3,0.2,0.2,0.3`). Priors on population size scaling factor η, the relative mutation rates of genes, and λ are all diffuse log-normals.

Three sets of simulations were done, Firstly, the program was run with no data in order to assess the prior. Various numbers of diploid and tetraploid species were tested. Secondly, a large number of simulations were run for the three scenarios shown in Figure 4. These were chosen to have similar topologies but different numbers of hybridizations. Different numbers of genes ($G = 1, 3, 9$), individuals per species ($N = 1, 3, 9$), and five mutation rates ($T = 5e{-}9, 1e{-}8, 2e{-}8, 4e{-}8, 8e{-}8$ mutations per site per generation) were tested. Note that since the root height is kept the same in terms of substitutions when the mutation rate is varied, different values of the mutation rate mean different numbers of generations from root to tip. When $T = 5e - 9$, then root height is $.012/5e - 9 = 2{,}400{,}000$ generations. When $T = 8e - 8$, the root height is 150,000 generations. Large values of T result in greater amounts of incomplete lineage sorting. The third set of simulations used a single scenario with 6 diploid species and 7 tetraploid species, as shown in Figure 5.

MCMC chains of up to 30M generations were used; for the $N = 9, G = 9$ case a few replicates failed to converge within 10M. Note that although there are only 5 species in scenarios D,E, and F, there are 8 diploid genomes and therefore $8 \times 9 \times 9 = 648$ sequences. These BEAST runs took around 6 hours each using one core on a desktop computer.

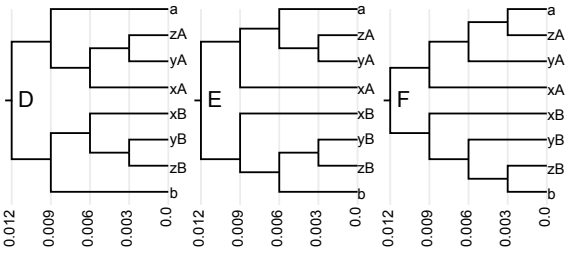

Figure 4: Scenarios D,E,F: the true MUL-trees. D has one hybridization; E has two; and F has three. Heights are in expected numbers of substitutions.

4.1 Empirical data

Silene.

5 Results

5.1 Distances for multi-labeled trees

In order to summarize the accuracy of the results it useful to have some definitions of distances for multi-labeled trees. We start with three variations of the Robinson-Foulds distance [19] for binary

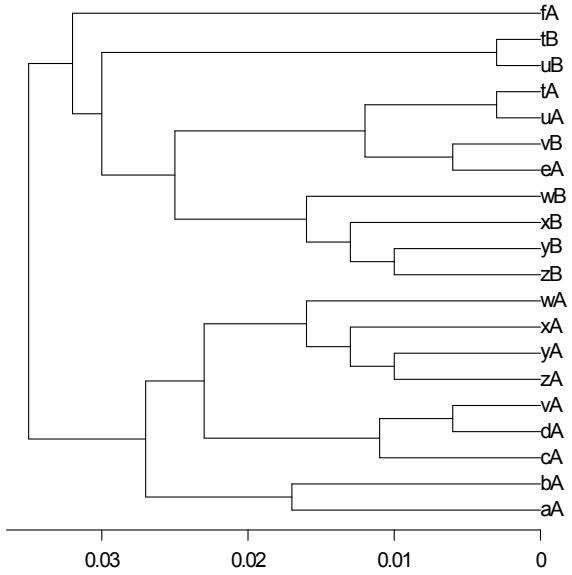

Figure 5: The true MUL-tree for a scenario with 6 diploid species labelled a,b,c,d,e,f and 7 tetraploid species comprising clades {t,u}, {v}, and {w,x,y,z}, arising from 3 hybridizations. Heights are in expected numbers of substitutions.

rooted trees. The first is D_{top} which purely topological; the other two, D_{miss} and D_{total}, include branch lengths. They are defined by the following algorithm. Given two binary rooted trees T_1 and T_2 with the same tip labels:

1. For each node i in T_j ($j \in \{1, 2\}$), find the clade C_{ji} and the length of the branch B_{ji} leading to C_{ji}.

2. Set $D_{top} = 0$, $D_{miss} = 0$ and $D_{total} = 0$.

3. For each clade C_{ji} which does not have a match in the other tree, add 1 to D_{top}, and add B_{ji} to both D_{miss} and D_{total}.

4. For each clade C_{ji} which does have a match C_{kl} (for some l and $k = 3-j$) in the other tree, add $|B_{ji} - C_{kl}|$ to D_{total}.

In order to extend this to multi-labeled trees M_1 and M_2, we follow [9] and define distances D_{top}, D_{miss} and D_{total} by considering all possible consistent relabellings of M_1 and M_2 and finding the minimum distance over all such relabellings. For the results here, this amounts to giving each pair of tips in M_1 an arbitrary labeling to distinguish them (say 'A' and 'B'), and then labeling each pair of tips in M_2 with either ('A','B') or ('B','A'). Thus, in order to evaluate a distance for a pair multi-labeled trees with m allotetraploids, 2^m ordinary Robinson-Foulds-type distances must be evaluated.

Note that $D_{miss} = 0$ implies that $D_{top} = 0$, and that $D_{total} \geq D_{miss}$ always. Also note that it is possible for different networks to correspond to the same multi-labeled tree; an example is shown in Figure 6. If there is an exact coincidence of node times of the two most recent diploid speciations in the second network, there is no way of distinguishing the two networks using the type of data used in this paper. If such exact coincidences do not occur, D_{miss} will be nonzero if a incorrect multi-labeled tree was inferred, although D_{top} may be zero.

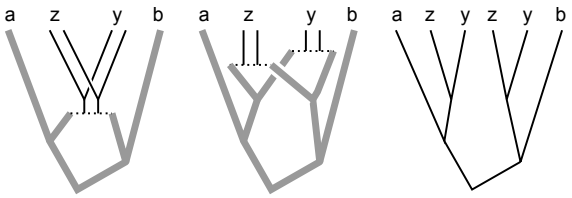

Figure 6: Two networks with the same multi-labeled tree.

5.2 Simulations

With no data, the prior given by equation 2 is sampled. Some examples of the marginal distribution for m are shown in Figure 7. It is fairly uniform over m. This is the case for the range of numbers of diploid and tetraploid species considered here, but not for much larger numbers species.

Scenarios D,E,F. Results are shown in Figure 8. As expected, accuracy increases with G, N, and decreases with T. In general, increasing G is more useful than increasing N. The dependence on T is less than might be expected, at least for $N = 3$ and $N = 9$. It is worth noting that estimates of m are often correct even when the topology is wrong (results not shown).

Scenario J.

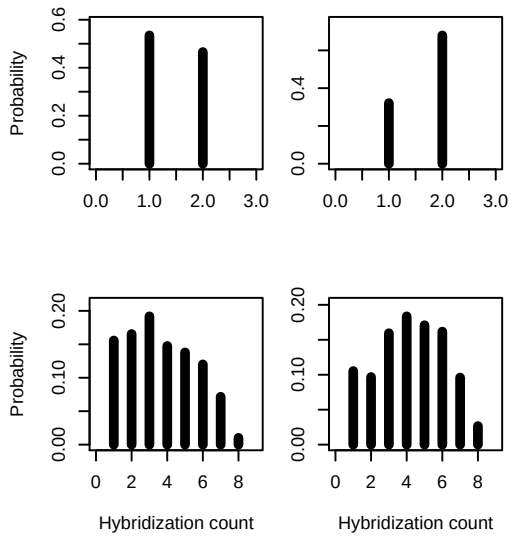

Figure 7: Estimates of the marginal distribution for the number of hybridizations m in the prior. There are 2 tetraploid species in the top row, and 8 in the bottom two. There are 2 diploid species in the left column and 8 in the right column.

5.3 Empirical data

Results for Silene data are shown in Figure 9. There is little signal in this data, and the posterior shows significant probability for both $h = 3$ (about 2/3) and $h = 2$ (about 1/3). In the latter case, Ss and Si arise from speciation after hybridization. In the figure, two MUL trees are shown, each conditional on thes possibilities.

6 Discussion

Mixing problems.

Heterozygosity. 6 assignments of 4 seqs from allotet indiv.

Hexaploids, etc.

Distances for multi-labeled trees.

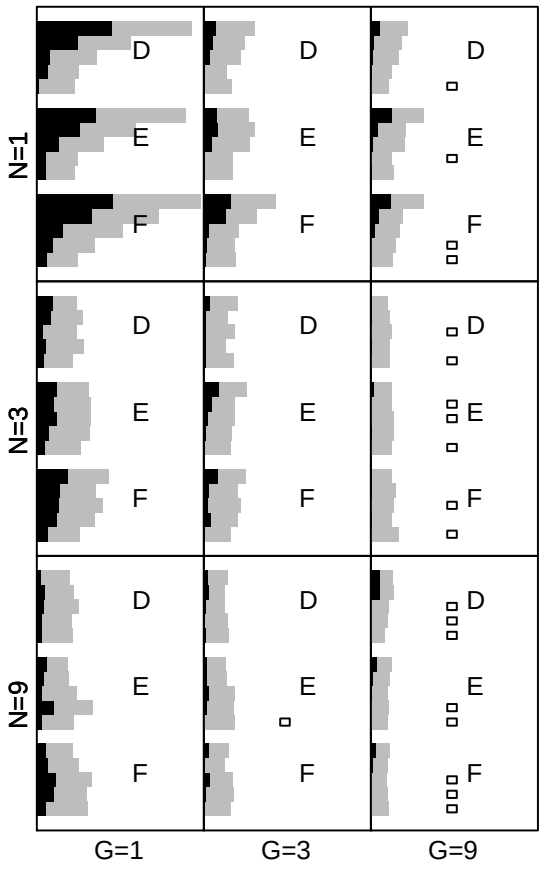

Figure 8: Results for scenarios D,E,F. The bars show mean values over 10 replicates for D_{total} values (gray) and D_{miss} values (black). Each group of five bars shows results for different mutation rates. Small open squares indicate cases $D_{top} = 0$ for all ten replicates.

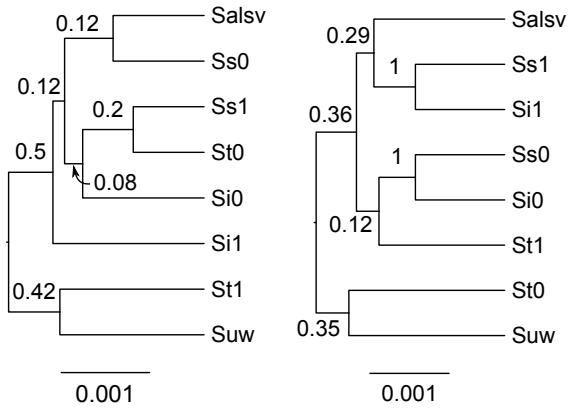

Figure 9: Results for Silene data. The MUL-tree on the left is conditonal on 3 hybridizations, and the one on the right is conditonal on 2 hybridizations.

References

[1] Zhi-Zhong Chen and Lusheng Wang. Hybridnet: a tool for constructing hybridization networks. *Bioinformatics*, 26:2912–2913, 2010.

[2] Zhi-Zhong Chen and Lusheng Wang. Algorithms for reticulate networks of multiple phylogenetic trees. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 9(2):372–384, March 2012.

[3] D. Gerard, H.L. Gibbs, and L. Kubatko. Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. *BMC Evolutionary Biology*, 11(1):291, 2011.

[4] T Gernhard. The conditioned reconstructed process. *J. Theo. Biol.*, 253:769–778, 2008.

[5] Peter J Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(1):711–732, 1995.

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

[7] K T Huber and V Moulton. Phylogenetic networks from multilabelled trees. *J Math Biol*, 52:613–632, 2006.

[8] K T Huber, B Oxelman, M Lott, and V Moulton. Reconstructing the evolutionary history of polyploids from multilabeled trees. *Mol Biol Evol*, 23:1784–1791, 2006.

[9] K T Huber, A Spillner, R Suchecki, and V Moulton. Metrics on multi-labelled trees: interrelationships and diameter bounds. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(4):1029–1040, 2011.

[10] G Jones, S Sagitov, and B Oxelman. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Syst. Biol.*, 2013.

[11] J.F.C. Kingman. *Poisson Processes*. Oxford University Press, 1993.

[12] L.S. Kubatko. Identifying hybridization events in the presence of coalescence via model selection. *Syst. Biol.*, 58(5):478–488, 2009.

[13] Martin Lott, Andreas Spillner, Katharina T. Huber, and Vincent Moulton. Padre: a package for analyzing and displaying reticulate evolution. *Bioinformatics*, 25(9):1199–1200, 2009.

[14] Martin Lott, Andreas Spillner, Katharina T Huber, Anna Petri, Bengt Oxelman, and Vincent Moulton. Inferring polyploid phylogenies from multiply-labeled gene trees. *BMC Evolutionary Biology*, 9:216, 2009.

[15] T Marcussen, K S Jakobsen, J Danihelka, H E Ballard, K Blaxland, A K Brysting, and B Oxelman. Inferring species networks from gene trees in high-polyploid North American and Hawaiian violets (Viola, Violaceae). *Systematic Biology*, 61(4):107–26, 2012.

[16] Bob Mau, Michael A Newton, and Bret Larget. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, 55:1–12, 1999.

[17] J. Pitman. *Combinatorial Stochastic Processes. Ecole d'été de Probabilités de St-Flour XXXII, Lecture Notes in Mathematics 1875*. Springer, 2006.

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

[19] D. Robinson and L. Foulds. Comparison of phylogenetic trees. 1981.

[20] Z Yang and B Rannala. Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences of U.S.A*, 107:9264–9269, 2010.