

# Allopolyploid network prior

Graham Jones

2011-09-17

## 1 Introduction

Calculating a prior probability for the network using a model of the process of speciation, extinction, and hybridization seems very difficult. The idea here is to suppose the diploid tree, develops as a standard birth-death process, and gives rise (somehow) to successful hybridizations. By ‘successful’ I mean that they leave at least one species at present time. After a successful hybridization event, the allotetraploid tree develops as a standard birth-death process.

This leaves two gaps:

What formula to use for the rate at which the diploid tree produces successful hybridizations?

What distribution to use for the ‘legs’ of an allotetraploid tree, that is, for the lineages leading from a branch of a diploid tree to the hybridization event?

### 1.1 Notation

$W$  is the network topology and node times. It can also be seen as a multiply labelled tree.

$\lambda$  is the speciation rate for both diploids and allotetraploids.

$\mu$  is the extinction rate for both diploids and allotetraploids.

$\gamma$  is the ratio of rate to speciation rate for both diploids and allotetraploids, so that the extinction rate  $\mu = \gamma\lambda$ .

$\beta$  is the ratio of hybridization rate for a pair of co-existing diploids to the speciation rate for a single diploid. So  $d$  diploids will produce an allotetraploid at rate  $(d(d-1)/2)\beta\lambda$  and a new diploid at rate  $d\beta\lambda$ .

$\delta = \lambda(1 - \gamma)$  is the diversification rate.

### 1.2 Rate for successful hybridizations

The number of diploids at a given time in the past is not known. If the extinction rate is high and the tree is old, it may have been a lot more in the past than at the present. If the extinction rate is low and the tree is young, it will have been growing roughly exponentially, and will have mostly been a lot less than at present. In theory, the age and extinction rate could be estimated, but that

would get very complicated. An approximation between possible extremes is a linear rate from the root of the diploid tree where there were two diploids to the present when there are  $d$ , say.  $d$  does not have to be the number of diploids in the analysis. It should be an estimate of the total number of extant diploids that are descendants of the mrca of the ones that are in the analysis. So the rate at which successful hybridizations occur is assumed to be

$$\alpha(t) = \beta\lambda \left( \frac{t}{t_0} + \frac{t_0 - t}{t_0} \right) \binom{d}{2} \quad (1)$$

where  $t_0$  is the time of the root of the diploid tree. This does not explicitly model the fact that a recent hybridization is more likely to be successful than an old one, because it has had less time to go extinct. However that seems likely to be a fairly small effect compared to the errors due to estimation of  $d$  and the linear approximation.

### 1.3 Distribution for the ‘legs’

It seems quite natural to place the ‘feet’ uniformly along any diploid branch in the diploid tree that is earlier than the hybridization event. In more detail, a point on a diploid branch is chosen for the first leg, and then a point for the second is chosen either on a diploid branch, or along the first leg. This does not model any evolutionary process of course: the legs are being added as an afterthought.

I also assume that there is no change in the probability of the network, regardless of how many legs there are, or how much total branch length is available to place their feet. This is probably the most dubious aspect of the scheme. However if this scheme is not used, some other arbitrary decision has to be made, for example about the distribution of the number of hybridization events. Note that by ignoring the legs in the probability of the network, the number of parameters remains constant as the number of hybridization events changes: a hybridization event is added and a speciation within an allotetraploid tree is subtracted.

I intend that this distribution will be directly sampled from during the MCMC process. There is a formula for the probability density for the node times and topology within homoploid trees, and for the hybridization times. MCMC moves affecting these will be made and the prior recalculated, but by assumption, the legs do not effect the probability of the network, so direct sampling is sufficient.

### 1.4 Formula for the network minus legs

$$\Pr(W|\lambda, \gamma, \beta) = \prod_{i=1}^r \Pr(T_i|h_i, \lambda, \gamma) \prod_{i=1}^r \Pr(h_i, \lambda, \beta, D) \Pr(D|\lambda, \gamma) \quad (2)$$

where  $D$  is the diploid tree,  $T_1, \dots, T_r$  are the allotetraploid trees, and  $h_1, \dots, h_r$  are the hybridization times.  $h_i$  is the origin (not root) of the allotetraploid tree  $T_i$ . The probabilities for the  $h_i$  are given by

$$\Pr(h_i, \lambda, \beta) = \alpha(h_i)$$

where  $\alpha()$  is as in equation (1)

DOES THIS NEED NORMALISING? What happens when  $\lambda$  changes?

## 2 The standard birth-death formulas

This is to record my understanding of the birth-death process as modelled in BEAST. Let

$$p(x) = \frac{\rho(\lambda - \mu)^2 e^{(\mu - \lambda)x}}{(\rho\lambda + (\lambda - \rho\lambda - \mu)e^{(\mu - \lambda)x})^2}.$$

and

$$q(x) = \frac{\rho(1 - e^{(\mu - \lambda)x})}{\rho\lambda + (\lambda - \rho\lambda - \mu)e^{(\mu - \lambda)x}}$$

Note that  $q'(x) = p(x)$ . I assume  $\rho = 1$  here. Then

$$p(x) = \frac{(\lambda - \mu)^2 e^{(\mu - \lambda)x}}{(\lambda - \mu e^{(\mu - \lambda)x})^2}.$$

and

$$q(x) = \frac{(1 - e^{(\mu - \lambda)x})}{\lambda - \mu e^{(\mu - \lambda)x}}$$

In terms of  $\delta$  and  $\gamma$ :

$$p(x) = \frac{(1 - \gamma)^2 e^{-\delta x}}{(1 - \gamma e^{-\delta x})^2}.$$

and

$$q(x) = \frac{(1 - \gamma)(1 - e^{-\delta x})}{\delta(1 - \gamma e^{-\delta x})}$$

and as logs:

$$\log(p(x)) = 2 \log(1 - \gamma) - \delta x - 2 \log(1 - \gamma e^{-\delta x}) \quad (3)$$

$$\log(q(x)) = \log(1 - \gamma) + \log(1 - e^{-\delta x}) - \log(\delta) - \log(1 - \gamma e^{-\delta x}) \quad (4)$$

Theorem 2.5 of [1] gives the densities for speciation times in the reconstructed tree, given the origin time  $t_0$ . The speciation times are iid, and the density for a speciation time at  $x$  is

$$p(x)\mathbb{1}(x \leq t_0)/q(t_0) \tag{5}$$

where  $\mathbb{1}$  is the indicator function.

For speciation times  $t_1, \dots, t_{s-1}$  in a tree with  $s$  tips,

$$\Pr(t_1, \dots, t_{s-1} | t_0) = q(t_0)^{-(s-1)} \prod_{j=1}^{s-1} p(t_j) \mathbb{1}(t_j \leq t_0)$$

with logarithm (assuming all  $t_j \leq t_0$ ) given by

$$-(s-1) \log(q(t_0)) + \sum_{j=1}^{s-1} \log(p(t_j)) \tag{6}$$

## 2.1 Allotetraploid trees

For the allotetraploid trees the origin is known to be the hybridization time, so I set  $t_0$  to be the hybridization time in equation (6) and use it for  $\Pr(T_i | h_i, \lambda, \gamma)$  in equation (2).

## 2.2 Diploid tree

For the diploid tree, the situation is different, because the origin time is unknown. This is the typical case and is implemented in BEAST. Below is working out what BEAST does.

Theorem 3.2 of [1] gives

$$\Pr(t_0 | s) = s\lambda^s q(t_0) p(t_0)$$

So

$$\Pr(t_0, t_1, \dots, t_{s-1} | s) = s\lambda^s p(t_0) \prod_{j=1}^{s-1} p(t_j) \mathbb{1}(t_j \leq t_0)$$

Check:

$$\int_0^t p(x) dx = q(t) - q(0) = q(t)$$

so the integral over region  $0 \leq t_1, \dots, t_{s-1} \leq t_0 \leq \infty$  is

$$X = \int_0^\infty s\lambda^s p(t_0)q(t_0)^{s-1} dt_0$$

From  $q'(x) = p(x)$ , we get  $(q(x)^s)' = sq(t_0)^{s-1}p(x)$  so

$$X = \lambda^s [q(x)^s]_0^\infty = \lambda^s (\lambda^{-s} - 0) = 1$$

End of check

Now for case where  $t_0$  is assumed to have an improper uniform prior on  $[0, \infty)$ , one could use  $\Pr(t_0, t_1, \dots, t_{s-1} | s)$  by introducing  $t_0$  as an extra parameter and sampling from it. But it can be integrated out. We want

$$\int_m^\infty s\lambda^s p(x) \prod_{j=1}^{s-1} p(t_j) dx = s\lambda^s \prod_{j=1}^{s-1} p(t_j) \int_m^\infty p(x) dx$$

where  $x = t_0$  and  $m = \max\{t_1, \dots, t_{s-1}\}$ .

$$\int_m^\infty p(x) dx = \int_m^\infty \frac{(\lambda - \mu)^2 e^{(\mu-\lambda)x}}{(\lambda - \mu e^{(\mu-\lambda)x})^2} dx$$

Set  $w = \lambda - \mu e^{(\mu-\lambda)x}$  and get

$$\begin{aligned} \frac{\lambda - \mu}{\mu} [-w^{-1}]_{\lambda - \mu e^{(\mu-\lambda)m}}^\lambda &= \\ \frac{\lambda - \mu}{\lambda} \frac{e^{(\mu-\lambda)m}}{\lambda - \mu e^{(\mu-\lambda)m}} & \end{aligned}$$

So we get

$$\frac{\lambda - \mu}{\lambda} \frac{e^{(\mu-\lambda)x}}{\lambda - \mu e^{(\mu-\lambda)m}} s\lambda^s p(x) \prod_{j=1}^{s-1} p(t_j) dx$$

Converting to logarithms in  $\gamma$  and  $\delta$  this is

$$\log(s) + (s-1)\log(\delta) - (s-2)\log(1-\gamma) - \delta m - \log(1-\gamma e^{-\delta m}) + \sum_{j=1}^{s-1} \log(p(t_j)) \quad (7)$$

This appears to be the formula used in BEAST.

```

private double loglikelihoodDiploidTree(Tree ditree) {
    int ntips = ditree.getExternalNodeCount();
    double z = Math.log(ntips);
    double delta = prior.getRate().getParameterValue(0);
    z += (ntips-1) * Math.log(delta);
    z -= (ntips-2) * Math.log(1.0-gamma);
    for (int i = 0; i < ditree.getInternalNodeCount(); i++) {
        double t = ditree.getNodeHeight(ditree.getInternalNode(i));
        if (ditree.isRoot(ditree.getInternalNode(i))) {
            z -= delta * t;
            z -= Math.log( 1.0 - gamma*(expdx(delta, t)) );
        }
        z += logp1(delta, t);
    }
    return z;
}

```

## References

- [1] “The conditioned reconstructed process”, *Journal of Theoretical Biology* Volume 253, Issue 4, 21 August 2008, Pages 769-778, doi:10.1016/j.jtbi.2008.04.005 (<http://dx.doi.org/10.1016/j.jtbi.2008.04.005>)