# Network prior based on trees with varying numbers of tips

Graham Jones

2011-11-07

## 1 Introduction

This note describes an idea for a prior for an allopolyploid network, for diploids and allotetraploids, in which the network is converted into a tree which has a number of tips depending on the number and sizes of the allotetraploid subtrees. The tree is not multiply labelled. It is intended for calculating a prior in which the distribution of the number of allotetraploidizations can be specified.

### The idea is not working. Seems worth recording, but no more.

For ordinary species trees, the distribution over labelled histories given by the Yule process is often used as a prior. As described in 2011-11-06-yule-density-on-trees, the density is

$$\frac{2^{n-1}}{(n-1)!}\lambda^{n-1}\exp[-\lambda(2t_1 + t_2 + \cdots + t_{n-1})] \tag{1}$$

where there are $n$ labelled tips and the ordered node heights are $t_1 \geq t_2 \geq \cdots \geq t_{n-1}$. The expression $(2t_1 + t_2 + \cdots + t_{n-1})$ is the total edge length of the tree and can be interpreted as the total length of time during which a speciation at rate $\lambda$ might occur. The idea is to do something similar for a network.

## 2 Details

Notation: see Figure 1. Suppose there are $d$ diploids and $m$ allotetraploids. Suppose there are $h$ hybridization events, and therefore $h$ allotetraploid subtrees $T_1, \ldots, T_h$. In the figure $d = 2, m = 6, h = 3$. For any tree $X$, let $|X|$ denote the number of tips in $X$, and $L(X)$ denote the total edge length of $X$. For those $i$ with $|T_i| \geq 2$, let $r_i$ be the root time of $T_i$. (Note that if $T_i$ has one tip, then the tip is also the root so has known time 0.) Let $b_i$ be the hybridization time for $T_i$, so that $b_i \geq r_i$. There is a 'diploid history' $D$ which has $d$ tips at present time 0, and $2h$ tips at times $b_i$, one pair of tips for each hybridization event. Let $s_i$ be the time of the most recent node in $D$ which is an ancestor of one of the tips at time $b_i$. More loosely, $s_i$ is the nearest node to the hybridization. Either one or both of the diploid species born at $s_i$ is involved in the hybridization at time $b_i$. For the case when $|T_i| \geq 3$, the root of $T_i$ has two child nodes. Let $c_i$ be the time of the most ancient child node, the one nearest the root.

The total edge length of a network is

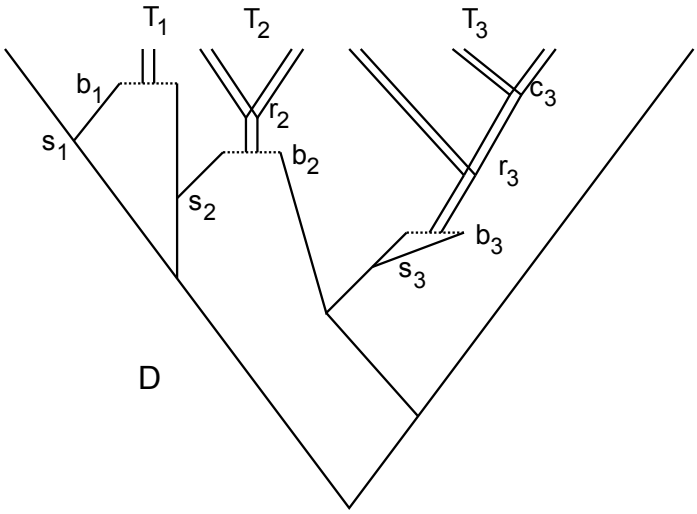


Figure 1: Network notation

$$L(D) + \sum_{i=1}^{h} L(T_i) + \sum_{i:|T_i|=1} b_i + \sum_{i:|T_i|>1} (b_i - r_i) \tag{2}$$

so the prior for a network most directly analagous to the Yule prior for a tree would be of form

$$K \ \exp\left[ -\lambda\left( L(D) + \sum_{i=1}^{h} L(T_i) + \sum_{i:|T_i|=1} b_i + \sum_{i:|T_i|>1} (b_i - r_i) \right) \right] \tag{3}$$

where $K$ is a normalization constant depending on $\lambda, |D|, |T_1|, \ldots, |T_h|)$. Unfortunately, this expression is not easy to integrate in the same way as a Yule prior for a tree. The problem is that the order of the node times in $D$ relative to the hybridization times $b_i$ affects which labelled histories are possible. If all the nodes in $D$ are earlier than all the $b_i$, then the tips of $D$ can join in any order, but otherwise there are constraints, and a smaller number of labelled histories are possible. This means that $K$ is difficult to calculate except for small cases.

If the terms in equation (2) involving $r_i$ and $b_i$ are ignored, it simplifies things somewhat. This approximation can be interpreted as replacing the total edge length of the network with what it would be if all hybridizations occurred as late as possible. But there is still the problem that the number of ways of labelling the tips depends on the topology.

Figure 2 illustrates the idea. This shows how the network in Figure 1 can be converted into a tree $S$ whose total edge length is that of the all the trees $D$ and $T_i$ together. In the case of a $T_i$ with just one tip, the two diploids are continued to the present. For allotetraploid subtrees with more than one tip, there will be a root, and two nodes (possibly tip nodes) just above the root. In this case one diploid is continued to each of these two nodes. This produces a binary tree without multiple labels. Note that no $r_i$ or $b_i$ appear in this expression: $2r_i$ appears in the edge length of $T_i$, but this is cancelled by $-2r_i$ in the edge length of $D$. The only node times are those internal to

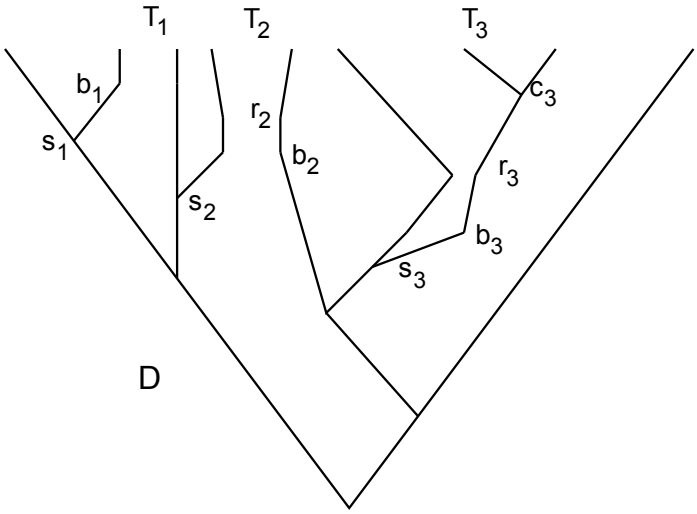


Figure 2: Network to tree

$D$, or internal to one of the $T_i$ but not the root of $T_i$. This process produces two tips for each $T_i$ with $|T_i| = 1$, and $|T_i|$ tips if $|T_i| \geq 2$. The number of tips is

$$n = d + \sum_{i=1}^{h} \min(2, |T_i|) \tag{4}$$

The catch is that the formula in equation 1 does not apply here. The number of ways of labelling the tips depends on the topology. The labelling within the $T_i$ is one problem – probably something could be done about that. But the labelling of the diploid history depends on the order of the nodes (in time) relative to the hybridization times. The more recent the nodes in the diploid history are, the more the topology is restricted.

# 3 Alternative?

It seems annoying that the number of tips varies with the sizes of the allotetraploid subtrees. However I can't see a good way around this. Figure 3 shows the sort of problem that arises. In the situation at the top, the roots of the allotetraploid subtrees can be joined to the diploid node which is at the earliest possible time of allotetraploidization. But in the lower case, this results in a gap. One could join it up somehow, but this either changes the total edge length, or the topology.

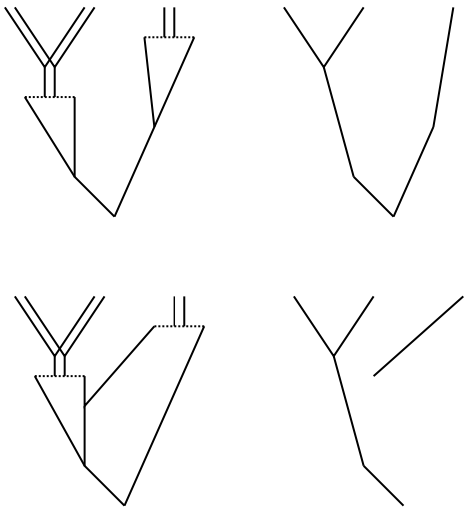

Figure 3: Alternative way of converting network to tree