

Yule process as prior for MUL tree

Graham Jones

2011-10-31

1 Introduction

Suppose there are m allotetraploids and d diploids. Their evolutionary history can be represented either as a network or as a multiply labelled tree (MUL-tree). I will call a particular sequence of evolutionary events (speciations, allopolyploidizations, extinctions) a ‘scenario’. I will call the network Φ and the MUL tree Υ .

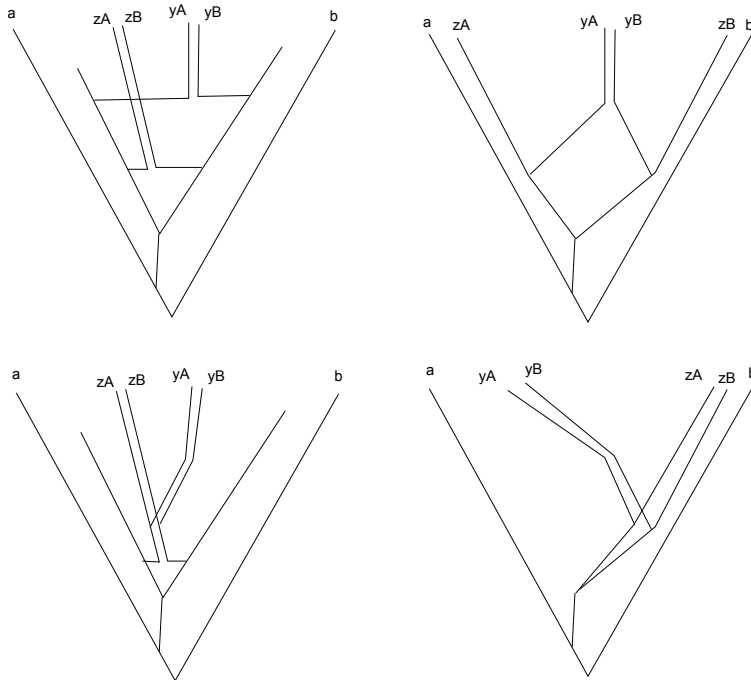


Figure 1: One or two hybridizations on left, same MUL tree on right.

1.1 PADRE

The number of allopolyploidizations is not determined by the topology of MUL-tree, but the minimum number of allopolyploidizations required can be found by the PADRE algorithm [2]. So far as I know, no algorithm current exists to find all the possible networks that could correspond to a given MUL-tree.

2 Visiting MUL-tree states in MCMC

One possibility is to use a MUL-tree as the model for the evolutionary history, and regard it as far as possible as a normal species tree. If a MUL-tree corresponds to a particular scenario with h ($1 \leq h \leq m$) allopolyploidizations, there will be h pairs of identical subtrees. These subtrees will have identical population sizes along branches. They will also have identical internal node times, of which there will be

$m - h$. This is something which does not happen in a normal species tree. To make the MUL-tree as much like a normal species tree as possible, one could allow the node times and population sizes to vary freely, ignoring the constraints.

The sequences cannot be assigned to tips a priori, but apart from that the model is like that for *BEAST. In particular there is no change in the number of parameters: the number of nodes and population sizes is fixed. MCMC moves are required for sequence assignment, but the MUL-tree only requires moves similar to those already implemented. This makes programming the model simpler than alternative options.

Once a MUL-tree has been estimated in this way, it may be possible to make further inferences about the evolutionary history, but these would be ad hoc inferences rather than statistical inferences.

In a Bayesian context, a prior on MUL-trees is needed in order to estimate a MUL-tree. The usual priors (based on the Yule or more general constant rate birth-death process) are not suitable. I will explain why and what might be done about it.

3 Prior for MUL-tree

3.1 approximate calculation

Here is an approximate calculation for the probability that $h = m$, in the case where $d = 0$ (no diploids), and given a prior based on the constant rate birth-death process (CRBD process). If $h = m$ it means the maximum possible number of allopolyploidizations has occurred, so that every allotetraploid has arisen from a separate allopolyploidization.

In order for h to be less than m , there must be at least one pair of cherries in the MUL-tree which have among their four tips only two labels, and furthermore, the labels must be like (a,b) and (a,b), not like (a,a) and (b,b). With $2m$ labels (two copies of $\{1, 2, \dots, m\}$) randomly assigned to tips, the probability that a particular pair of cherries contains just two particular labels is the probability of choosing 4 particular labels out of a total of $2m$. This is

$$\frac{4}{2m} \frac{3}{2m-1} \frac{2}{2m-2} \frac{1}{2m-3} = \frac{24}{2m(2m-1)(2m-2)(2m-3)}$$

There are $\binom{m}{2}$ pairs of labels, so the probability of two cherries containing just two labels is

$$\frac{m(m-1)}{2} \frac{24}{2m(2m-1)(2m-2)(2m-3)} = \frac{3}{(2m-1)(2m-3)}$$

If the two cherries contain just two labels, there is a 2/3 chance that the two cherries have identical labels (ie like (a,b) and (a,b), not like (a,a) and (b,b)). So the probability of two cherries having identical labels is

$$\frac{2}{(2m-1)(2m-3)}$$

If there are r cherries, there are $\binom{r}{2}$ pairs of cherries. If we pretend that the probabilities for two cherries having identical labels are independent, the probability of no pair of cherries having identical labels is

$$\left(1 - \frac{2}{(2m-1)(2m-3)}\right)^{r(r-1)/2}$$

For a CRBD process with n tips, the expected number of cherries is $n/3$ [1]. Making a second approximation, the probability of no pair of cherries having identical labels is

$$\left(1 - \frac{2}{(2m-1)(2m-3)}\right)^{(m/3)(2m/3-1)}$$

As $m \rightarrow \infty$ this tends to $\exp(-1/9) \approx 0.895$.

I haven't calculated the probability for just one allopolyploidization ($h = 1$) under the CRBD model, but under the Proportional to Distinguishable Arrangements (PDA) model, it is

$$\frac{(2m-3)!!}{(4m-3)!!} = \frac{1}{(2m-1)(2m+1)\dots(4m-3)}$$

This is 1/315 for $m = 3$ and about 4e-15 for $m = 10$.

3.2 Simulation

I wrote a simulation to generate random tree topologies under the CRBD process, with labels assigned from two copies of $\{1, 2, \dots, m\}$ and counted the number of times that there was no pair of cherries with identical labels.

m	Estimate of Pr(no matching cherries)
2	.778
3	.848
4	.869
5	.876
10	.887
20	.890
100	.894

So it seems the approximation is very good, even for small m . Note that 'no matching cherries' implies that $h = m$, but h could be m even if there are matching cherries. The probability for just one allopolyploidization $\Pr(h = 1)$ is too small to estimate in this way.

3.3 Consequences

Using a 'normal' prior for a MUL-tree therefore means assuming that the maximum number of allopolyploidizations has occurred with a probability .8 to .9. It also means assuming a tiny probability for just one allopolyploidization. This is just a prior – sufficient data will overwhelm it – but it seems a very poor choice.

3.4 Modification

One could use the PADRE algorithm [2] to find the minimum number of allopolyploidizations of each MUL-tree visited by the MCMC chain and give greater weight those with fewer required allopolyploidizations. The algorithm seems to be fast enough, but it is not clear how one should do the weighting. The issues are:

- The PADRE algorithm finds the minimum number of allopolyploidizations, but it is not known what other evolutionary histories would produce the same MUL tree. So it is not known what is being assumed about the probabilities of these.
- There is no known formula for the minimum number of allopolyploidizations. The above calculations and simulations give a rough idea.

References

- [1] Andy McKenzie and Mike Steel, *Distributions of cherries for two models of trees*, *Mathematical Biosciences* 253, 164 (2000) 81–92

- [2] K. T. Huber, B. Oxelman, M. Lott and V. Moulton, *Reconstructing the evolutionary history of polyploids from multi-labelled trees*, Mol Biol Evol 2006 23: 1784–1791
- [3] Bob Mau, Michael A. Newton, Bret Larget, *Bayesian Phylogenetic Inference via Markov Chain Monte Carlo Methods*, Biometrics, Vol. 55, No. 1 (Mar., 1999), pp. 1–12
- [4] Peter J Green, *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*, Biometrika (1995), Vol. 82, 4, No. 1 pp. 711–32