Trees with legs: phylogenetic analysis of allopolyploids

Graham Jones

Talk given 2011-12-08, this version 2012-01-10

1 Introduction

I have been working on a project with Bengt and Serik involving allopolyploids. I'll explain what those are in a moment.

I have been working on this since April, and most of my time has been spent extending BEAST (which is a program for phylogenetics) to deal with allopolyploids. My model is most similar to that of *BEAST.

Here's the outline.

Outline

Species trees and allopolyploid networks

Gene trees

The Bayesian approach

Progress

Future

2 Species trees and networks

2.1 Triangle of U

This is an example from the cabbage family, taken from Wikipedia.

The three species at the corners existed as separate diploid species. Each pair hybridized to create a new species of tetraploid Brassica.

The important thing to notice is that the genomes are added in the new species, not merged together. n is the number of chromosomes found in the pollen or ovule. So 8 here plus 9 here makes 17 and so on.

2.2 What is an allopolyploid?

Allopolyploids are polyploids that arise by hybridization between species. (Coyne an Orr, p323)

That would be simple if anyone knew what a species was! More seriously, this is quote from a paper about *BEAST.

Please bear in mind that 'species' ... is not necessarily the same as a taxonomic rank, but designates any group of individuals that after some 'divergence' time, have no history of breeding with individuals outside that group. A species tree defines barriers for gene flow, and so the term is a catch all for taxonomic rank, subspecies, or any diverging 'population structure'. (Heled and Drummond, 2010)

For allopolyploids, tree becomes network.

Genomic allopolyploids carry entire chromosome sets from two or more parental species. The sets of chromosomes are so different that they do not pair at meiosis.

In **Segmental allopolyploids** some chromosomes do and some don't.

To summarise what I've just said about species and allopolyploids: it is gene flow that matters, since that is what we can amke inferences about.

2.3 Disomic segregation

This shows the way that segregation happens in genomic allopolyploids. It traces the history of an individual. The red and green chromosomes stay separate.

The model I am developing is appropriate for genes which have evolved like this.

2.4 Genetic distance

Looking at things at things fron a larger scale. Here are two diploids which hybridiz=ed at some point to form an allotetraploid.

If a gene is acting according to the model, the genetic distances between the four sequences should look something like this.

A big difference between those coming from different parental species, and a small difference within the diploid genomes.

Spelling it out: The alleles which come from the **same** parental species (and therefore the same diploid genome) will usually be very similar, perhaps homozygous. Ones which come from the **different** parental species will usually be quite different, since they have been evolving separately for a long time.

I will assume that the differences between alleles from the same parental species can be ignored. The mathematical model is that one allele is chosen at random from one of the diploid genomes, and one from the other diploid genomes. So just two sequences from each individual plant for each gene.

2.5 Allotetraploids - no slide

I will draw the line at allotetraploids in this talk. So there may be some diploids and some of their ancestors may have hybridized to form allotetraploids, but no further hybridization occurs.

2.6 Representations for allopolyploid networks

Actual evolutionary events. Includes things that there is no or little information about in the genes - extinct species and hybridization time.

Trees with legs. Similar to 'phylogenetic networks' that are often used to display information about hybridization. Easiest for changing the number and attachments of the tetraploid subtrees.

Multiply labelled trees. Neatest maths. Easiest for fitting gene trees into.

The computer code I have written uses both these two.

The case on the right, omitting the diploids, gives a simple situation, that I will use a particular situation as an example in this talk...

2.7 A simple situation

So here they are again, without the diploids. The lines indicate the hybridization time.

On the right, the same situation but drawn as tubes.

There was originally one diploid, which speciated here.

They may speciate again, but at some point two descendants of the original diploid hybridize - here, and after that, all the diploids go extinct.

Finally the new allotetraploid species splits into two.

There is therefore no data from any diploids, just from the two allotetraploids.

Now consider how the gene trees can fit into this....

3 Gene trees

The available data for making inferences is multiple sequence alignments of genes from the extant species.

From these gene trees can be inferred. There are two main problems with this: **incomplete lineage sorting and sequence ambiguity.**

3.1 Incomplete lineage sorting - usual

Looking back from the present, two genes can coalesce only if they belong to the same population.

This is the standard sort of diagram that illustrates the problem of incomplete lineage sorting.

3.2 Incomplete lineage sorting

Of course it gets more complex with allopolyploids...

For allopolyploids, two genes can coalesce only if they belong to chromosomes which recombine at meiosis. Assuming we have genes that don't do that unless they come from the same parental species, that means they can't coalesce until you trace them back to a single diploid ancestor. The red ones can't coalesce with green ones until they get right to the bottom of this network.

3.3 Incomplete lineage sorting - no colour!

The other main problem is that when sequences are obtained from an allopolyploid, it is not possible to say which is which. They aren't conveniently coloured red and green for example.

3.4 Sequence alignment

The data that we do have looks like this. A multiple sequence alignment.

I have used different colours and different numbers to indicate different individual plants.

'x' and 'y' indicate different species.

The new thing that we have with allotetraploids are labels for the two sequences from each individual. I have used letters A and B for these labels. So we know that A and B represent different parental species, but we **don't** know which is which. So I am using A and B to distinguish them, but A and B don't mean anything in themselves.

3.5 Sequence ambiguity 1

Putting this together you get something like this.

This is the gene tree sitting inside the species network.

This is the same tree untangled, and with labels representing the sequences as I showed before.

The upper magenta line is when genes sampled in different species a and b can coalesce.

The lower magenta line is when genes with different parental species can coalesce.

Since the labels for the sequences within each individual can all be flipped, there are 2 times 2 times 2 times 2 equals 16 possibilities. But there are really only 8 different ones, since we don't care about whether the first sequence is called A or B. What does matter is how the labellings realte to one another.

3.6 Sequence ambiguity 2

I have shown the 8 possibilities here. I have fixed the order of the first A and B, the red ones, and listed the ways the others could be assigned.

The condition from the parental species means that if two symbols appear on either side of the vertical line, they cannot coalesce until the bottom.

For example, if we look at the first one, there is a red B one side and a green A on the other. But the gene tree shows those coalesced very recently. That means that this assignment of sequences is incompatible with the species network.

Only assignment 4 works in this case.

[Given a compatible gene tree, and population sizes, it is possible to calculate a probability for the coalesences, in a similar way to starBEAST.]

4 The Bayesian approach

So far I have been assuming a particular evolutionary history. Now I turn to the problem of inferring the evolutionary history from the data.

One approach to the problem is to assume the gene trees are true, or nearly true, or that most of them are nearly true, and then try to fit a network around them. Mostly these approaches have only used the topology of the gene trees, when there may be useful information in the times of coalescences.

I am trying another approach - a statistical approach, and particularly Bayesian approach. This says...

4.1 Co-estimate everything

Write down an expression for the likelihood of the whole thing, given parameters:

- The identities of the sequences
- The topology and node times of the species network
- The topology and node times of all the gene trees
- The populations of all the lineages (along all branches)
- The substitution matrix, rates of evolution along branches, site rate variation, and a few other things, for all the gene trees

Assume a prior for all the parameters.

Calculate.

4.2 What *BEAST does

This is essentially the formula that *BEAST uses.

You multiply over different genes, or if you like different alignments.

S is a tree with parameters being: topology, node times, populations at each node.

4.3 New model

Now make it more complicated to deal with networks. Remember that the network can be seen as a set of trees wih legs, or as a multiply-labelled tree. I'll use W to represent it, and call W a network or a multiply-labelled tree as seems appropriate.

W is a now a multiply-labelled tree. It still has parameters: topology, node times, populations at each node. But now some of the labels at the tips are ambiguous.

$P(g_i|W)$ is similar to what *BEAST does, when W is viewed as a multiply-labelled tree.

$P(d_i|g_i, \gamma_i)$ can be thought about in various ways. I'll say more about this later.

4.4 Priors

Need priors. Won't say anything about that today.

4.5 Calculation: New MCMC moves

For now, I restrict to diploids and tetraploids. Five types of MCMC proposals needed.

1. Changes within the tetraploid subtrees.

2. Changes within diploid tree, dealing with cases where branches that lead to a tetraploid change time or disappear.

3. Changes to the way a tetraploid subtree joins diploid tree (times and branches).

4. Changes to the number of tetraploid subtrees.

- the hardest, because it involves changes in the number of parameters and therefore needs reversible jumps.

5. Changes to sequence assignments.

5 Progress

It seemed best to extend *BEAST.

5.1 Current state

I have implemented a model in BEAST for the case of two diploids and a single hybridization event. This means I didn't have to implement the hardest MCMC move (4 above, changes to the number of tetraploid subtrees) and changes to the diploid tree are very simple - just the root time.

I haven't got any real data yet.

5.2 Simulations

Essential to test on simulations.

However, I have written simulation code. This makes a multiply labelled tree, then generates gene trees using a coalescent model, then uses Seq-Gen (written by Andrew Rambaut) to generate sequences. The BEAST XML from the sequences. Then run BEAST. Then TreeAnnotator.

5.3 One allotetraploid tree

The first case I chose to tackle is one with two diploids and a single tetraploid.

a and b are diploids. z is a tetraploid.

These show three scenarios - three ways in which z might have arisen from a and b. There are a couple more which are mirror images.

5.4 Results

Here are some preliminary results.

The results are better when root is more ancient. Not surprising, it is an easier problem, because deep coalescences are less common.

More genes (G) help. There is more data.

Adding extra individuals per species hardly helps at all. This changes if the branches that need resolving are more recent. Then extra individuals per species *do* help.

6 Future

6.1 Allow more than one hybridization

The main thing to add is the ability to deal with more than one hybridization.

In particular, rjMCMC moves to change the number of hybridizations by merging and splitting the tetraploid subtrees.

7 equations

$$P(S, g|D) \propto P(S)P(g)\Big(\prod_{i=1}^{n} P(d_i|g_i)P(g_i|S)\Big) \tag{1}$$

$$P(W, g, \gamma|D) \propto P(W)P(g)P(\gamma)\Big(\prod_{i=1}^{n} P(d_i|g_i, \gamma_i)P(g_i|W)\Big) \tag{2}$$