Phylogenetic analysis for allopolyploids

Graham Jones

2011-11-30

1 Introduction

Allopolyploidy is an important mechanism for generating new species, which is particularly prominent in plants [[8], [2], [9] are refs from New Phytologist, not sure which are relevant to allo vs auto. Also [1].] Allopolyploids are produced by hybridization between two species. Hybridization presents a challenge to phylogenetic analysis since the usual tree is replaced by a network. This article explores the feasibility of making statistical inferences about the evolutionary history of allopolyploids using simulations of some simple scenarios and a novel model implemented in the BEAST [3] framework.

The main restrictions made here are that there has been a single hybridization event between two diploids, and that the resulting hybrid is a *genomic allopolyploid*, in which the two diploid genomes (from the two parental diploid species) do not recombine with one another at meiosis because the chromosomes in the two parental species were too diverged by the time the hybrid formed. This leaves two major problems to deal with in the phylogenetic analysis. Firstly, when the organisms are sequenced, it is not possible to assign sequences to their parental diploid species. Thus, although the sequences can be seen as the result of the evolution of diploid genomes, there is an ambiguity in the labelling of the sequences which is not normally present. Secondly, the issue of incomplete lineage sorting cannot be ignored, because the hybridizations can be recent.

There are various ways of viewing the evolutionary history of a set of genomes of this kind. Figure 1 shows three ways of viewing the same evolutionary events: as a network or as a multiply labelled tree (MUL tree). In the top row, hybridization and extinction events are explicitly represented. The second row shows the network view. The networks can be understood as a collection of homoploid 'trees with legs' in which trees of higher ploidy are connected by their legs to those with lower ploidy. In the MUL tree view, there is a binary tree, but some of the tips have the same labels since they correspond to the same species.

The data available for inferring the species history consists of molecular sequences sampled from individuals belonging to species. One approach, pursued in [5], [6], and [7], is to estimate the gene trees first, and then search for the MUL tree that best accomodates them. Here the approach is the typically Bayesian one of 'co-estimating everything'. The network node times and topology, the assignment of sequences to parental diploid species, and the node times and toplogies of all the gene trees are all allowed to vary, and an MCMC algorithm is used to sample the posterior distribution. The approach is similar to that of *BEAST [4] but the sequence ambiguity is new.

Figure 2 shows an example of sequence ambiguity. [probably use a different example.]

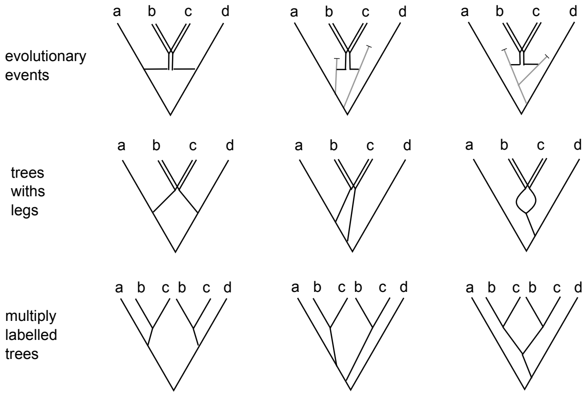

Figure 1: Representations

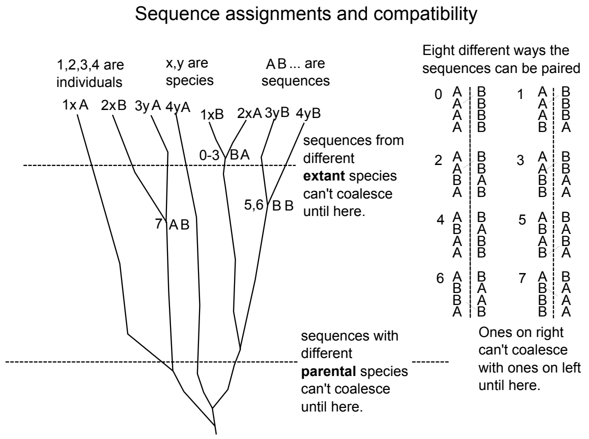
Sequence assignments and compatibility
1,2,3,4 are individuals
x,y are species
AB... are sequences
1xA 2xB 3yA 4yA 1xB 2xA 3yB 4yB
0-3 BA
7 AB
5,6 BB
sequences from different **extant** species can't coalesce until here.
sequences with different **parental** species can't coalesce until here.
Eight different ways the sequences can be paired
0 AAAA BBBB
1 AAAB BBBA
2 AABA BBAB
3 AABB BBAA
4 ABAA BABB
5 ABAB BABA
6 ABBA BAAB
7 ABBB BAAA
Ones on right can't coalesce with ones on left until here.

Figure 2: Sequence ambiguity

2 Model and methods

It is assumed that at some point in the past a diploid species speciated to form two diploid species which both continued to the present. The initial speciation forms the root of the network or MUL tree. At some later time a hybridization took place between these two diploid species or their extinct relatives, forming a tetraploid in which the two parental diploid genomes continue to evolve without recombining with one another at meiosis. [probably add speciation of the tetraploid.]

A multi-species coalescent model, similar to that of *BEAST [4] is used within species. As in *BEAST, it is assumed that there is free recombination between genes, but no recombination within genes. Also, the term 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" [4].

2.1 The formula

The network topology and node times is denoted by W. It can also be seen as a multiply labelled tree. The population size parameters are denoted by θ. There is one value at each tip, one at the rootward end of each branch, and one just after the hybridization event. The (hyper-)parameters for the topology and node times W are denoted by λ. Here λ is a speciation rate, but in a more sophisticated model it could include an extinction rate and a hybridization rate. η is the population mean, appearing in a hyperprior for θ. For the gene trees, there are the following parameters. G is the number of gene trees. τ_i is the i'th gene tree topology and node times. α_i is all the other parameters belonging to the i'th gene tree: parameters for site rate heterogeneity, substitution model, branch rate model, root model. g_i is (τ_i, α_i), that is, all the parameters for the i'th gene tree. γ_i is the permutations of sequences within individuals for the i'th gene. d_i is the sequence data for the i'th gene. $\tau = (\tau_1, \ldots \tau_G)$, and similarly for α, g, γ, d. The posterior probability is then given by

$$\begin{aligned}
\Pr(W, \theta, g, \gamma | d) \quad \propto \quad & \Pr(W|\lambda) \Pr(\lambda) \times & (1) \\
& \Pr(\theta|\eta) \Pr(\eta) \times & (2) \\
& \Pr(\gamma) \times & (3) \\
& \prod_{i=1}^{G} \Pr(\tau_i | W, \theta, \gamma_i) \times & (4) \\
& \prod_{i=1}^{G} \Pr(d_i | g_i) & (5)
\end{aligned}$$

The five terms in this expression will now be described in detail.

1. $\Pr(W|\lambda) \Pr(\lambda)$ is the network prior: the probability of W before seeing any molecular data. [More to say, still thinking about what is best here, but for the simple scenarios investigated and with the assumption of a single hybridization this is not very hard or critical.]

2. $\Pr(\theta|\eta) \Pr(\eta)$ is the population prior. The priors used here were similar to those typically used by *BEAST. An independent gamma distribution is assumed for each population size. The

shape parameter is 4 for the populations at the tips, and 2 for the rest. An improper prior proportional to $1/\eta$ was assumed for the population mean η.

3. $\Pr(\gamma)$ is the permutation prior. This is assumed to be uniform here, so can be omitted.

4. $\Pr(\tau_i|W, \theta, \gamma_i)$ is the probability of τ_i, when permuted by γ_i, fitting into the network W with populations determined by θ. Note that this probability does not depend on α_i.

   One can think of the γ_i as permuting the sequence data d_i, that is swapping pairs of rows in a data matrix for the i'th gene, where the pairs have been sequenced from the same individual. Thus one could write

   $$\Pr(d_i|g_i, \gamma_i)\Pr(g_i|W, \theta)$$

   as

   $$\Pr(\gamma_i(d_i)|g_i)\Pr(g_i|W)$$

   This doesn't work well in implementation in BEAST, where it seems simplest to regard γ_i as permuting the tip labels of the gene tree as indicated in Figure 2. The sequences attached to a tip don't change, nor does the gene tree topology. Instead, the way in which the sequences are assigned to tips in the multiply labelled tree W is changed.

   Apart from this extra complexity due to the permutations γ_i, the value of $\Pr(\tau_i|W, \theta, \gamma_i)$ is similar to that used in *BEAST. The populations are assumed to very linearly along branches in the MUL tree, between the nodes where the population parameters occur. However, the population is allowed to change discontinuously at hybridization. Within a given branch, or within the two parts of a branch before and after hybridization, the formula is

   [to be copied from [4], with refs going back from there.]

5. $\Pr(d_i|g_i) = \Pr(d_i|\tau_i, \alpha_i)$ is the 'Felsenstein likelihood' of the data for the i'th gene given the i'th gene tree.

2.2 Scenarios

Three scenarios, as shown in Figure 3 were used. Each scenario represents a 'true' MUL-tree. Heights are in expected substitutions per site. All allopolyploidizations occur at 0.01. Population sizes are numbers of gene copies within diploid populations, or numbers of gene copies with the same diploid parent, for allotetraploid populations. If the population size is S, the probability of coalescence between a pair of gene copies is $1/S$ per generation.

All genes have length 500. Population sizes are 100,000 at tips, and at rootward ends of branches, and 200,000 at tipward ends of internal branches and at the root.

These three scenarios were each tested with the number of genes G equal to 1,3,10, the number of individuals N per species equal to 1,3, and the mutation rates T set to 4e-8 and 8e-8. This is a total of 12 possibilities per scenario.

The T values are in expected substitutions per site per generation. All scenarios have a root height of 0.04. Changing T while keeping this height fixed changes the number of generations the tree

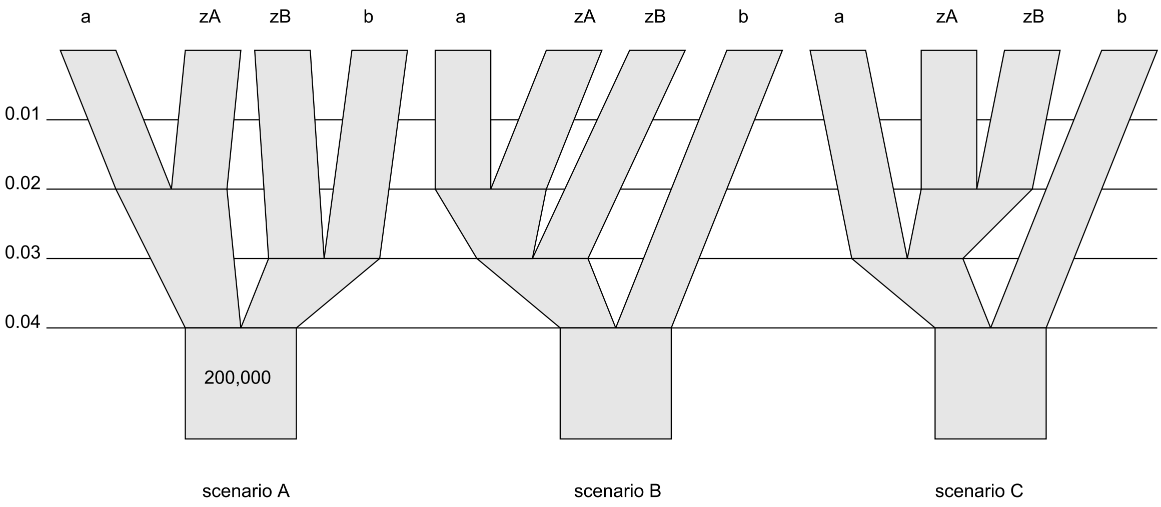

Figure 3: Scenarios

represents. T=4e-8 means 0.04/4e-8 = 1e6 or a million generations root to tip. T=8e-8 means 0.04/8e-8 = 5e5 or half a million generations root to tip.

2.3 Other parts of the model

Strict clock branch rates assumed in simulation and BEAST.

HKY substitution model assumed in simulation and BEAST. In simulations, the Seq-Gen parameters are kappa=3, frequencies .3 A and T, .2 C and G. (Seq-Gen called with `-t3.0 -f0.3,0.2,0.2,0.3`. These are estimated in BEAST.

No site rate heterogeneity assumed within genes in simulation and BEAST.

Genes have the same mutation rate in simulations. Rates are estimated in BEAST.

2.4 Implementation of simulations

[To do. How much detail?]

2.5 Implementation of allopolyploid model in BEAST

[To do. How much detail?]

3 Results

[There will be more simulated results. Plus real data.]

For each scenario, and each set of 12 (G, N, T) values, 20 replicates were simulated and run for one million generations in BEAST, a total of 720 BEAST runs. This took about ten hours on one processor per scenario, the bulk of time running BEAST. MUL-trees were sampled every 1000 generations, and the first 200 samples (of 1001) discarded as burn-in. The tables below show the results as number of times the correct topology of the MUL-tree was recovered as the concensus tree.

Scenario A	Generations root to tip	
G, N	1,000,000	500,000
1, 1	16 / 20	12 / 20
1, 3	19 / 20	12 / 20
3, 1	18 / 20	17 / 20
3, 3	20 / 20	15 / 20
10, 1	20 / 20	19 / 20
10, 3	20 / 20	19 / 20

Scenario B	Generations root to tip	
G, N	1,000,000	500,000
1, 1	12 / 20	13 / 20
1, 3	16 / 20	9 / 20
3, 1	19 / 20	18 / 20
3, 3	20 / 20	13 / 20
10, 1	20 / 20	20 / 20
10, 3	18 / 20	17 / 20

Scenario C	Generations root to tip	
G, N	1,000,000	500,000
1, 1	14 / 20	9 / 20
1, 3	15 / 20	12 / 20
3, 1	18 / 20	18 / 20
3, 3	20 / 20	17 / 20
10, 1	20 / 20	20 / 20
10, 3	20 / 20	20 / 20

3.1 Discussion of simulated results

[Mostly to be done]

I think the general conclusions will be:

- The length of the branches to be resolved has a major impact on the difficultly of the problem, as you would expect. The key quantity is the ratio of the length of the branch measured in generations to the population size. If the is ratio is less than one, the problem is hard, and will require lots of data.
- More genes improve the results.
- More individuals can improve the results, but if the branches that need to be resolved are ancient, or if there is a population bottleneck, the addition of more individuals is not much help.

3.2 Empirical data

[To be done]

3.3 Discussion of empirical data

[To be done]

References

[1] Coyne, J.A., Orr, H.A.: Speciation. Sinauer Associates (2004)

[2] Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Carlson, J.E., Arugumanathan, K., Barakat, A., et al, V.A.A.: Widespread genome duplications throughout the history of flowering plants. Genome Research **16**, 738–749 (2006)

[3] Drummond, A.J., Rambaut, A.: BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology **7**(214) (2007)

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

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

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

[7] Lott, M., Spillner, A., Huber, K.T., et al: Inferring polyploid phylogenies from multiply-labeled gene trees. Bmc Evolutionary Biolog **9**, ?–? (2009). DOI 10.1186/1471-2148-9-216

[8] Wendel, J.F., Doyle, J.J.: Polyploidy and evolution in plants. In: R.J. Henry (ed.) Plant diversity and evolution: genotypic and phenotypic variation in higher plants, pp. 97–117. CABI Publishing, Cambridge, MA, USA (2005)

[9] Wood, T.E., Takebayashi, N., Barker, M.S., Mayrose, I., Greenspoon, P.B., Rieseberg, L.H.: The frequency of polyploid speciation in vascular plants. Proceedings of the National Academy of Sciences, USA **106**, 13,875–13,879 (2009)