Outline of idea for dealing with migration by an approximation

Graham Jones

2016-10-15, October 31, 2016

This is a work in progress...

This note is aimed at species tree estimation using the multispecies coalescent, in the presence of some migration. It is based on an approximation which will break down if there is a lot of migration.

'Migration' is used here to refer to gene flow between species (usually introgression but not restricted to that). A migration event occurs when an allele comes from a parent in another species. An 'embedding' of a gene tree specifies which species tree branch each coalescence belongs to, together with migration events, which specify the times along gene tree branches at which an allele moved between species tree branches, and which species tree branch is the destination. We always describe events going back in time from the present, so alleles have parents to which they 'go', and a destination branch refers to an earlier time than a source branch. This is because coalesences are easier to model this way, and is the same convention as IMa2 (Hey, 2010).

There is no upper limit on the number of possible migration events, and it is difficult to model the situation in full. IMa2 requires that the true population phylogeny (equivalent to species tree here) is known. The method of Dalquen et al. (2016) can estimate the species tree, but is currently restricted to at most 3 species and 3 sequences per locus, although it can handle very large numbers of loci.

Here we assume a simple model for migration. In IMa2 and Dalquen et al. (2016), migration rate parameters are defined between pairs of contemporaneous species tree branches. Here a single migration rate parameter is assumed which applies to all species tree branches in the same way. We also approximate the posterior by ignoring most of the 'unlikely' embeddings. If the migration rate is high, some of the ignored embeddings will be quite likely and the approximation will break down.

Clearly this is a rather crude approximation to reality. The hope is that it is better to account for migration crudely than to ignore migration altogether. Leaché et al. (2014) showed migration causes problems for species tree inference using the multispecies coalescent when migration is present but ignored. The model here is not suitable for a detailed investigation of migration between particular species.

General formula

Suppose G_j are the parameters specifying the topology and node heights for the jth locus and S is similar parameters for the species tree. Suppose E_j are parameters which specify how the jth gene tree is to be embedded into the species tree. The E_j provide enough information to do this in a unique way whatever the values of G_j and S are.

We want to decompose the joint probability density

$$\Pr(G_1, \ldots, G_J, E_1, \ldots, E_J, S).$$

where J is the number of loci. We assume that the the gene trees and their embeddings, that is, (G_j, E_j) for different j, are independent given the species tree, so the density is

$$\left(\prod_j \Pr(G_j, E_j|S)\right) P(S) = \left(\prod_j \Pr(G_j|S, E_j)P(E_j)\right) P(S).$$

We now describe a specific model of this general type.

The density for an embedded gene tree

We need the density for an embedded gene tree, namely $\Pr(G_j|S, E_j)$. This is essentially the model of Hey and Nielsen (Hey and Nielsen, 2004, 2007; Hey, 2010), where the multispecies coalescent model is combined with a model for migaration. Most of the notation here comes from Jones (2016). In branch b, we have effective population size N_b, mutation rate μ_b, and $\theta_b = N_b\mu_b$. The ploidy of gene j is p_j. We introduce a single migration parameter m. The rate at which an allele leaves one species (going back in time, as always) is assumed to be $(mN_b\mu_b)^{-1} = (m\theta_b)^{-1}$ when measured in substitution units. The coalescent part of the formula is similar to that of Jones (2016), but as in Hey and Nielsen's model, we must now account for the fact that lineages can enter and leave a species tree branch via migration.

For locus j, there are k'_{jb} coalescences in species tree branch b, and k''_{jb} events where a lineage leaves branch b, and k'''_{jb} events where a lineage arrives in branch b, totalling $k_{jb} = k'_{jb} + k''_{jb} + k'''_{jb}$. These divide the branch into $k_{jb} + 1$ intervals, of lengths c_{jbi} $(0 \le i \le k_{jb})$. Suppose that there are n_{jbi} lineages during the ith interval. During this interval, a lineage leaves branch b at rate $n_{jbi}(m\theta_b)^{-1}$.

$$\begin{aligned}
f_G(G|\Theta) &= \prod_j \prod_b (p_j\theta_b)^{-k'_{jb}} (m\theta_b)^{-k''_{jb}} \exp\left(-\sum_{i=0}^{k_{jb}} c_{jbi}\left[\binom{n_{jbi}}{2}(p_j\theta_b)^{-1} + n_{jbi}(m\theta_b)^{-1}\right]\right) \\
&= \prod_b \prod_j p_j^{-k'_{jb}} m^{-k''_{jb}} \theta_b^{-k'_{jb}-k''_{jb}} \exp\left(-\sum_{i=0}^{k_{jb}} c_{jbi}\left[\binom{n_{jbi}}{2}p_j^{-1} + n_{jbi}m^{-1}\right]\theta_b^{-1}\right) \\
&= \prod_b r_b \theta_b^{-q_b} \exp\left(-\gamma_b\theta_b^{-1}\right)
\end{aligned}$$

where

$$q_b = \sum_j k'_{jb} + k''_{jb}, \qquad r_b = \prod_j p_j^{-k'_{jb}} m^{-k''_{jb}}, \qquad \text{and} \quad \gamma_b = \sum_j \sum_{i=0}^{k_{jb}} c_{jbi}\left[\binom{n_{jbi}}{2}p_j^{-1} + n_{jbi}m^{-1}\right]. \tag{1}$$

As in Jones (2016), the population size parameters can be integrated out if a suitable form for the prior is assumed.

How the gene tree is embedded

The posterior arising from this model is hard to sample from. There is no upper limit on the number of migration events, and even if this is limited, there can be a huge number of ways in which each gene tree can be embedded into a species tree (even when the gene tree and species tree are fixed). We now describe a scheme for exploring a

region of parameter space which includes most of the probability content when m is small enough. How small m needs to be for this to be a good approximation remains to be seen.

Embeddings are restricted by applying the following rules:

1. there is at most one migration in a single gene tree branch
2. at most one of the child branches of a gene tree node contains a migration
3. there are no more migrations than needed (in a sense described below)

The parameters E_j consist of three values for each internal node of the jth gene tree. The internal nodes are indexed by i, using an ordering based on an ordering of the tip labels. There are two boolean parameters ξ_{ji} and η_{ji} and a continuous parameter $h_{ji} \in [0, 1]$. Here ξ_{ji} and η_{ji} each identify one of its child nodes (the smallest or largest in the ordering). The parameter ξ_{ji} specifies which of the node's child branches is capable of migrating; the other is assumed unable to migrate. The parameter η_{ji} specifies which of the node's child branches to use when choosing a destination species branch for an introgression. Finally, h_{ji} specifies the time at which a migration occurs, as a fraction along the gene tree branch.

The first rule above is straightforward. The definition of ξ_{ji} enforces the second rule. The third rule is applied recursively from the tips. Suppose x is the jth node of the jth gene tree, and suppose both child nodes of x have been assigned to branches in the species tree. If it is possible to assign x to a species tree branch without an migration in either child branch of x, then this is done. Otherwise x is assigned using the species tree branch to which its non-migrating child has been assigned: it will be the same branch, or an ancestor of that branch, depending on the height of x. The height of the migration is fixed by h_{ji}. The migrating child branch of x starts in the species tree branch that the migrating child has been assigned. It stays in this branch, or an ancestor of it, until the migration height. It will then migrate to the same species tree branch as x, or a descendant of it. If there is more than one descendant of the species tree branch of x at this height, values from η_j are used to choose one.

Most nodes x (small migration rates) will be assigned to species tree branches without involving ξ_{ji}, η_{ji} or h_{ji} at all. If a migration is needed, η_{ji} may or may not be needed; sometimes further values of η_{jk} ($k \neq i$) may be needed.

TODO. Discussion of $\Pr(E_j)$.The prior must avoid any dependence on the node order, so iid or at least exchangeable(?). It should be as 'uniform' as possible though it's not clear exactly what that means, especially η. TODO.

Properties of the embedding scheme

Different embeddings of the same gene tree in the same species tree are obtained by changing ξ_{ji}, η_{ji}, and h_{ji} during the MCMC sampling. Figure 1 shows some examples. Case (a) is simple. No migrations are needed to embed the gene tree, so embeddings with one or more migrations are ignored. Case (b) requires one migration, and an embedding with two migrations in the same branch is ignored. Case (c) requires two migrations. The embedding on the left is ignored since it has two sister branches with migrations. The embedding on the right is one of four embeddings that is considered.

Proposition *Given any set of particular values for ξ, η and h, and the rules above, any gene tree can be embedded in any species tree. For any G_j and S, the set of embeddings as ξ, η and h vary include some ones with a minimal number of migrations.*

Proof: First claim proved by recursion starting at the tips. The gene tree tips are assigned to species tree branches; then look at a gene tree node whose children have been assigned. TODO

For second claim, suppose it is false and consider the set M of minimal embeddings (those with a minimal number of migrating branches). Call a node both of whose child branches migrate a 'double node'. Thus every member of M has at least one double node. Now restrict attention to the subset $\bar{M}$ of M of embeddings which have as few as possible double nodes. Finally, choose an embedding B from $\bar{M}$ so that a double node x is as near to the root as possible.

If x is the root, it can be moved into the same branch as one of its children, or an ancestor of that branch, and one migration can be removed, contradicting the definition of M. If x is not the root, it can be again moved into the same branch as one of its children, but now the branch between x and its parent may need to become migrating. If the sister branch to x is already migrating, we have an embedding with the same number of migrations, but a double node closer to the root than x, contradicting the definiton of B. If the sister branch to x is not migrating, we have an embedding with fewer double nodes than B, contradicting the definition of $\bar{M}$. End of proof.

The method does not consider every embedding which has a minimal number of migrations (eg Figure 1c). Some embeddings which are considered are not minimal. Eg gene tree ((a1,b1),b2) with b2 and b2 in same species, and a species tree with large root height. If (a1,b1) is assigned to the branch a1, two migrations will be used, but it is possible to use only one.

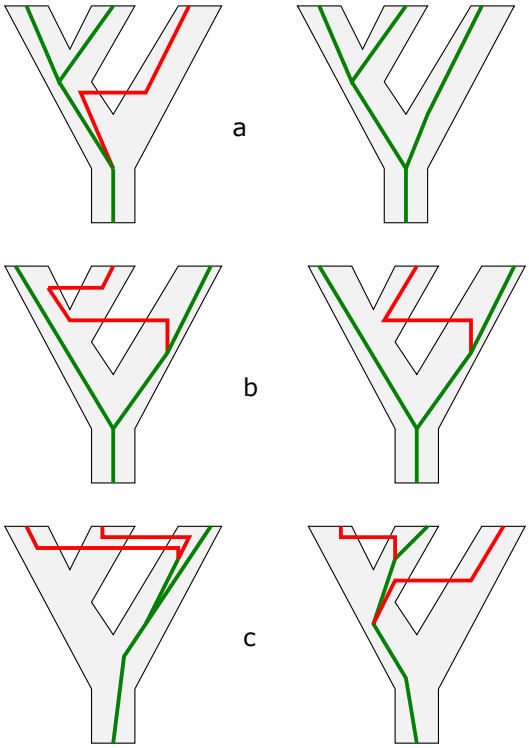

Figure 1: Some examples of embeddings for three gene trees in a, b, c. Branches which introgress are in red. On the left are embeddings that are ignored. On the right is an embedding which is considered.

Discussion

The rate at which alleles leave a species is smaller when the population is large, which makes sense since there are more parents within the species to 'choose' from. But in this model, no account is taken of the number of contemporaneous species, nor their population sizes. It may be possible to use more sophisticated models with m varying over species tree branches.

I think that by including more information in E_j, it would be to consider every embedding which has a minimal number of migrations. However, this appears difficult to implement, and especially difficult to implement efficiently. It is not clear how much difference this issue makes to species tree estimation, since some minimal embeddings are always considered, and the ignored embeddings appear to be fairly rare, occurring only when two or more migrations are needed near to one another.

Extinction. If there are extinct species, migration can result in unusually deep coalescences. TODO

I know of two simulation programs MCcoal and CoMuS2. TODO

References

Daniel A. Dalquen, Tianqi Zhu, and Ziheng Yang. Maximum likelihood implementation of an isolation-with-migration model for three species. *Systematic Biology*, 00:00–00, 2016.

J Hey. Isolation with migration models for more than two populations. *Mol Biol Evol*, 27:905–920, 2010.

J Hey and R Nielsen. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of drosophila pseudoobscura and d. persimilis. *Genetics*, 167:747–760, 2004.

J. Hey and R. Nielsen. Integration within the felsenstein equation for improved markov chain monte carlo methods in population genetics. *Proc Natl Acad Sci*, 104:2785–2790, 2007.

Graham Jones. Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *Journal of Mathematical Biology*, 2016. doi: 10.1007/s00285-016-1034-0. URL `http://link.springer.com/article/10.1007/s00285-016-1034-0`.

A D Leaché, R B Harris, B Rannala, and Z Yang. The influence of gene flow on species tree estimation: A simulation study. *Systematic Biology*, 63(1):17–30, 2014.