Phylogenetic Inference using RevBayes

Gene tree-Species tree reconstruction

Overview: Gene tree-species tree models

Ever since Zuckerkandl and Pauling, people have recognised that phylogenies reconstructed from homologous gene sequences could differ from species phylogenies. As molecular sequences accumulated, the link between gene trees and species trees started to be modelled. The first models were based on parsimony, and aimed for instance at reconciling a gene tree with a species tree by minimizing the number of events of gene duplication and gene loss. In the past dozen years, probabilistic models have been proposed to reconstruct gene trees and species trees in a rigorous statistical framework. Models and algorithms have quickly grown in complexity, to model biological processes with increasing realism, to accommodate several processes at the same time, or to handle genome-scale data sets. In this overview we will not detail these models, and we invite the interested reader to take a look at recent reviews (e.g. (**?**)).

Processes of discord

There are several reasons why a gene tree may differ from a species tree. Of course, a gene tree may differ from the species tree just because a mistake was made during the analysis of the gene sequences, at any point in a pipeline going from the sequencing itself to the tree reconstruction. Such a mistake would produce an incorrect gene tree. Here we do not mean this kind of discord, but rather discord that has comes from a real biological process that builds true gene histories that differ from true species histories. These processes include gene duplication, gene loss, gene transfer (used loosely here to also include reticulation, hybridization between species), and incomplete lineage sorting (1). Incomplete lineage sorting will be discussed in more details in the following subsection.

Fig. 1 suggests that for all processes the gene tree can be seen as the product of a branching process operating inside the species tree. As a consequence, all processes are modelled as some type of birth-death process. For duplication/loss models, birth correspond to gene duplication events, and death to gene loss events. Transfers can be added to the model by introducing another type of birth, with a child lineage appearing in another branch of the species tree. Incomplete lineage sorting is also modelled with a birth-death type of model, the coalescent. All these models can be made heterogeneous, by allowing different sets of parameters for different branches of the species tree. This is useful to model differences in rates of duplication, loss or transfer among species, or to model different effective population sizes in a species tree. In RevBayes so far only models of incomplete lineage sorting have been implemented (models of duplication and loss and transfer will soon be added). Thanks to RevBayes modular design, there is quite a lot of flexibility in specifying the model, for instance by allowing different parameters to different branches of the species tree, and the gene tree-species tree model could be combined to other types of models, for instance models of trait evolution.

Modelling incomplete lineage sorting: the multispecies coalescent

Incomplete lineage sorting is a population-level process. In a species, at a given time, there are several alleles for a given locus in the genome. These alleles have their own history, they diverged from each other at various times in the past. This history can differ from the species history, because several alleles can persist through a speciation event, and because, short of selective effects, the sorting of alleles during a speciation event is random and can result in a tree that differs from the species tree (Fig. 1d). In all cases, incongruence between the gene tree and the species tree occurs when alleles persist over the course

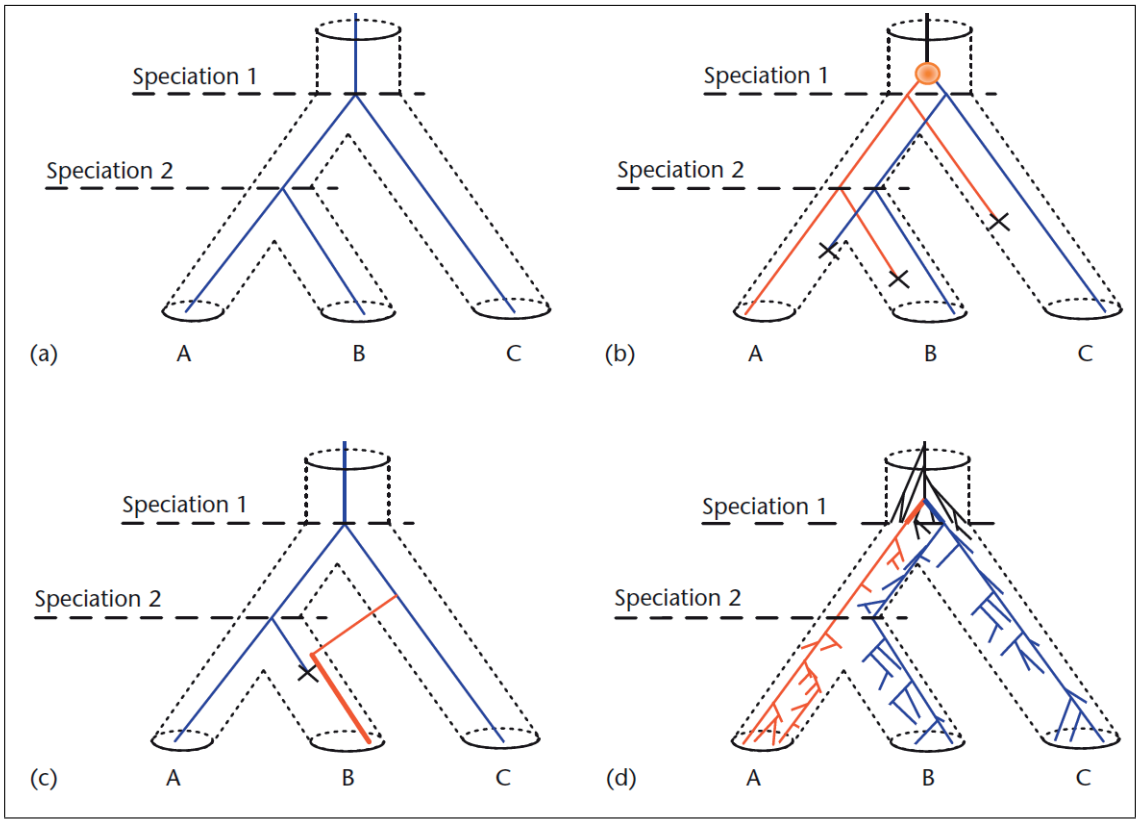

Figure 1: The processes of discord. The species tree is represented as a tubular structure. Gene trees are blue and red lines running along the species trees. a) A gene tree that perfectly matches the species tree. b) The gene tree and the species tree differ because of gene duplications and losses. c) The gene tree and the species tree differ because of gene transfer and gene loss. d) The gene tree and the species tree differ because of incomplete lineage sorting. [Replicated from Fig. 2 in **?**.]

of several speciation events. When reconstructing a gene tree, one therefore gets the history of the alleles that have been sampled (at best), not necessarily the history of the species.

In 2003, Rannala and Yang proposed a powerful way to model the sorting of alleles along a phylogeny of several species (**?**), the multispecies coalescent (Fig. 2). This model is at the origin of most model-based approaches to reconstruct gene and species trees (**??**). The multispecies coalescent appropriately models the evolution of a population of alleles along a species tree. Along the species tree, it allows different branch lengths, in units of time, and also allows different effective population sizes. Computing the probability of a gene tree given a species tree and other parameters is quite easy. Bascially it works by cutting the gene tree into independent species-specific subtrees, computing probabilities for each of those subtrees, and combining them all at the end to get the probability of the gene tree according to the multispecies coalescent, given the current parameter values. Cutting the gene tree into species-specific subtrees is quite easy, because we can use the dates of speciation events to know what's before and after speciation events. The resulting subtrees are represented with the grey boxes in Fig. 2. In this figure, each subtree corresponds to one particular population, either extant or ancestral. Inside each subtree, given its length, the effective population size, and dates of coalescence (alleles splitting), the coalescent model provides simple formulas for computing the probability of the gene subtree given other parameters. These subtree probabilities are then multiplied to get the gene tree probability given current parameter values.

Two parameters associated to branches of the species tree have a direct impact on the expected amount of gene tree-species tree incongruence:

- **Time between speciations.** The more a branch length increases, the more the pool of alleles is expected to change. Alleles are therefore less likely to persist for several speciation events if the branches between these speciation events are long.
- **Effective population size between speciations.** In populations with small effective population sizes, chance events can cause large shifts in allele frequencies, and possibly disappearance of alleles, irrespective of the fitness of this allele. In large populations, because an allele is likely carried by a large number of individuals, its disappearance is less likely, the population of alleles is more stable. Alleles are therefore less likely to persist for several speciation events if the branches between these speciation events are characterised by small effective population sizes.

Overall, larger amounts of gene tree-species tree incongruence are expected in phylogenies characterised by short branches with large population sizes. A corollary of that is that larger amounts of gene tree-gene tree incongruence are expected as well. To measure the susceptibility of species phylogenies to generate incomplete lineage sorting, the concept of *coalescent time units* has been introduced. Coalescent time units are obtained when branch length λ is divided by effective population size N_e. As a consequence, in a species tree whose branches are expressed in coalescent time units, a branch length of 1 *coalescent time unit* means a branch length of N_e *generations*. Once branch lengths on the species tree are measured in coalescent time units, it becomes easy to spot species trees that generate a lot of incongruence: those are short trees.

First RevBayes exercise: simulating gene trees under the multispecies coalescent

1. Open RevBayes
2. Let's simulate a species tree with 10 taxa, 10 gene trees, 5 alleles per species (feel free to change these values).

```
n_species <- 10
n_genes <- 10
n_alleles <- 5
```

3. We simulate a species tree topology according to a birth-death process with arbitrary parameter values (similar to **?**):

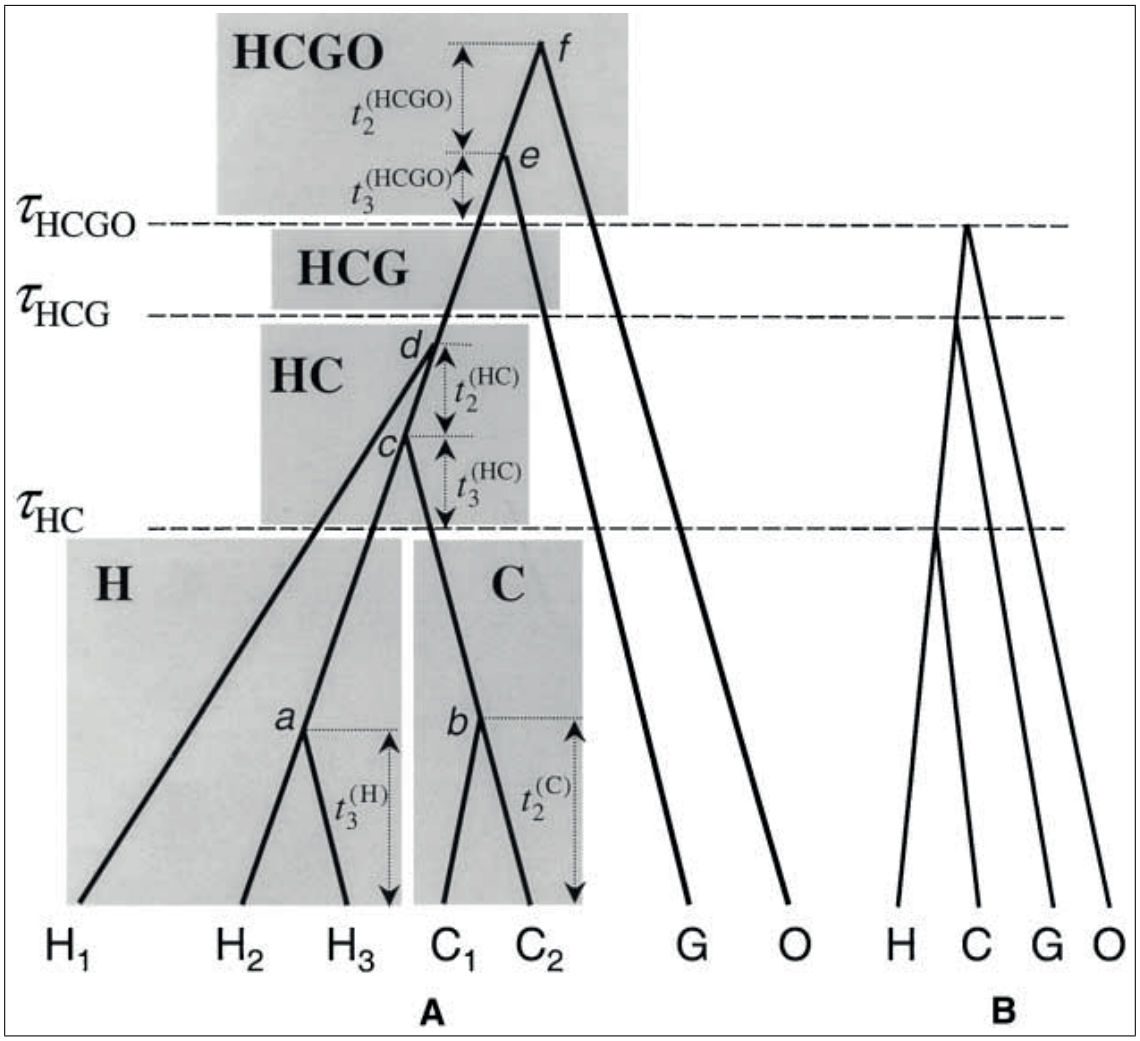

Figure 2: The multispecies coalescent. A) A gene tree, including 3 human alleles, 2 Chimp alleles, one Gorilla allele, and one Orang-outan allele. τ parameters are speciation times, t parameters are divergence time in the gene tree, the grey squares represent the ancestral populations, with their respeciive sizes. B) The corresponding species tree. In this model, the speciation times define minimal boundaries for allele divergence times. [Replicated from Fig. 1 in ?.]

```
speciation ~ exponential(10.0)
extinction ~ exponential(10.0)
tree_height ~ unif(0,1.0)
speciation.setValue(2)
extinction.setValue(0.3)
tree_height.setValue(0.8)
speciesTree ~ cBDP(lambda=speciation, mu=extinction, origin=
   tree_height, nTaxa=n_species, names=s_names)
```

4. Then we can use the multispecies coalescent model to generate gene trees. These can be examined

using Figtree or NJplot or any other tree viewer, but we can also directly compute symmetric differences between these from RevBayes. First, we simulate a set of gene trees, using a single effective population size for all branches, and after having constructed a map between species names and gene names:

```
# Build the mapping between sequence names and species names.
for (i in 1:n_species) {
        for (j in 1:n_alleles) {
                taxa[(i-1)*n_alleles+j] <- taxon(taxonName=s_names[i
                   ]+"_"+j, speciesName=s_names[i])
        }
}
# Set the effective population size
Ne ~ gamma(shape=0.1,rate=0.1)
Ne.setValue(0.004)
# Simulate gene trees
for (i in 1:n_genes) {
   # The gene tree from the multispecies coalescent process
   # Note that if Ne had been a vector of effective population sizes
      instead of a single value,
   # allowing 1 parameter per branch of the species tree, the same
      line would work.
   geneTrees[i] ~ dnConstPopMultispCoal(speciesTree=speciesTree, Ne=
      Ne, taxa=taxa)
}
```

5. We can compute symmetric differences between all these gene trees. The symmetric difference between two trees is the total number of partitions found in one tree but not in the other one. In our case, the maximal difference is as follows:

```
maxDiff <- 2 * (n_species*n_alleles - 2)
```

6. That will give us a reference for comparing with the values we get on our gene trees. We can build a function for computing all pairwise symmetric differences between our gene trees, and getting the mean.

```
function RealPos symDiffVector ( Real[] vec ) {
        ndiff <- 1
        for (k in 1:(n_genes-1)) {
                for (j in (k+1):n_genes) {
                        diff[ndiff]<-symDiff (geneTrees[k], geneTrees
                           [j])
                        ndiff <- ndiff+1
                }
        }
        return (mean(diff))
}
#We can then use this function on our gene trees:
symDiffVector(geneTrees)
```

7. Now to get a sense of how population size and branch lengths alter the gene tree distribution, we can relaunch the multispecies coalescent simulation (step 4) and look at the resulting gene trees after having rescaled the species tree or changed the effective population size. To do these little changes:

```
# Changing Ne:
Ne.setValue(0.08)
#Rescaling the species tree:
speciesTree.rescale(0.1)
```

Using the functions above, it is possible to look at the species tree in coalescent time units (which is very convenient). How would you do that?

Do these observations seem coherent with the multispecies coalescent presentation above?

Alternatives to the multispecies coalescent model

Strengths and weaknesses of the multispecies coalescent

The multispecies coalescent model is an elegant model. As we have seen above, we can easily simulate data under this model. Inference using this model, combined with a model of sequence evolution, is also possible, and when it works, is very informative. Not only can we get a dated species tree and dated gene trees from the multispecies coalescent, we can also get extant and ancestral effective population sizes (**??**). However inferring many parameters is difficult, and convergence can be difficult to reach with such models. In particular, the strong interdependence that exists between the gene trees and the species tree makes it easy for algorithms to fall into local maxima. As a consequence, there are ongoing efforts to develop methods for which inference would be easier, albeit at the cost of approximations and simplifications.

Alternatives to the multispecies coalescent

An easy way to simplify the problem is to consider that parts of it are already solved. For instance, several approaches assume that rooted gene trees are available. The problem then becomes markedly easier, but inference is then highly dependent on the quality of the input gene trees. Often, such methods also make other simplifying assumptions, and e.g. do not try to estimate separately time and effective population sizes, but instead directly work with coalescent time units. These methods usually are much faster than the multispecies coalescent, should be more robust against local maxima, but are less ambitious about the amount of information they can get from the data, and are can be sensitive to the quality of the input gene trees.

Another approach is to use methods that mathematically bypass estimating gene trees altogether. To our knowledge, there are three such approaches: SNAPP(**?**), POMO(**?**), and XX(**?**). They differ in the way the algorithms work, but they are all based on the same idea, which is integrating out gene trees. To achieve that they extend the model of sequence evolution, which usually models substitution events, to also model population-level processes. In RevBayes, so far only the POMO model has been implemented and will be discussed in the following part.

The POMO model

POMO models allele frequency changes along with mutations with a single transition matrix. It extends the usual 4×4 DNA substitution matrix to incorporate polymorphic states. In doing so, it makes a first important approximation: it only considers biallelic states. For instance, it considers sites at which either an A or a C is found in a population, but it won't consider sites at which 3 different states, e.g. A, C, T, are observed in a population. Then it makes a second approximation, which introduces a single virtual population size in lieu of the branchwise effective population sizes. This virtual population size is not inferred, but is fixed to some low number. In practice, **?** consider that a virtual population size of 10 individuals should produce good results. This virtual population size directly constrains the types of polymorphic states that can be considered. With 10 individuals for instance, only frequencies such as $(100\%A)$, $(10\%A, 90\%C)$, $(20\%A, 80\%C)$, $(30\%A, 70\%C)$, ..., $(90\%A, 10\%C)$, $(100\%C)$ can be considered. The POMO matrix models transitions among all these states, polymorphic ones as well as monomorphic ones, and has a size that depends on the virtual population size. For instance, with a virtual population size of 10 individuals, the POMO square matrix contains 58 rows: 4 monomorphic states, plus 6 types of biallelic states (AC, AG, AT, CG, CT, GT) times 9 frequencies. Additional assumptions of the POMO model include total independence among sites (no linkage among sites), and absence of mutations in biallelic states: transitions among biallelic states or from biallelic states to monoallelic states only occur through population-level changes in allele frequencies, not through mutation of one allele into another. Mutations only occur to transit from a monoallelic state to a biallelic state.

The POMO model therefore makes several approximations to avoid estimating gene trees. Fewer parameters need to be estimated, as neither gene trees nor population sizes are estimated, but other parameters can be introduced into the model. For instance **?** estimate 4 base fitness parameters, which they use to model GC-biased gene conversion, which tends to increase the GC content of recombining sequences. In the RevBayes implementation of POMO, base fitnesses can be estimated as well.

Inference using the multispecies coalescent and the POMO model

In this section we will perform inference using both the multispecies coalescent and the POMO model. Depending on your machine and on the size of the data, successful inference may take some time. If this tutorial is done in a classroom environment, it may be wise to convene with a friend that one tries the multispecies coalescent model while the other one tries the POMO model, and that results will be shared.

Batch Mode

If you wish to run this exercise in batch mode, the files are provided for you.

You can carry out these batch commands by providing the file name when you execute the **rb** binary in your unix terminal (this will overwrite all of your existing run files).

- **$ rb full_analysis.Rev**

Useful Links

- RevBayes: https://github.com/revbayes/code
- Tree Thinkers: http://treethinkers.org

Questions about this tutorial can be directed to:

- Bastien Boussau (email: boussau@gmail.com)
- Sebastian Höhna (email: sebastian.hoehna@gmail.com)

Version dated: August 24, 2014