Hybrid-coal manual

Sha (Joe) Zhu

February 24, 2016

1 Introduction

In phylogenetic studies, trees are used for describing evolutionary histories. In particular, a species tree presents population divergences, and a gene tree indicates the times when genes started to differentiate within populations. Even though speciation is driven by gene mutations, using a single gene tree to infer the species tree is not ideal. Often, the inconsistency among gene trees and species trees makes describing the relationships between and among species very difficult. Common causes of the conflict include gene duplication, horizontal gene transfer, incomplete lineage sorting, and hybridization [5, 7]. If speciation events occur close together, it is likely that some gene copies remain the same after species divergence. This inconsistency between gene trees and species trees is referred as incomplete lineage sorting.

Hybridization refers to interbreeding between species. Offspring who carry genes from both parental species then reproduce and form a new species. For closely related species, however, both lineage sorting and hybridization are likely to occur, e.g. an avian genus *Manacus* [1] or the New Zealand alpine cicadas [2]. Here, I present research on the probabilistic modelling of coalescence with lineage sorting in hybridized species. In these models, the relationships among species are represented by a network rather than a tree, while relationships at the gene level are still represented by trees.

`hybrid-coal` has been developed to calculate the gene tree probabilities within a species network.

2 Download and installation

`hybrid-coal` can be downloaded from https://github.com/hybridLambda/hybrid-coal. Extract the source code by executing the following command:

```
tar -xf hybrid-coal-VERSION.tar.gz.
```

It is fairly standard to compile `hybrid-coal` on UNIX-like systems. In the directory `hybrid-coal-VERSION`, execute the following command:

```
$./bootstrap
$make
```

3 Notation

3.1 Input/output formats

The input file for `hybrid-coal` is a character string that describes the relationships among species. Standard Newick format is used for inputting species trees and outputting gene trees, e.g.:

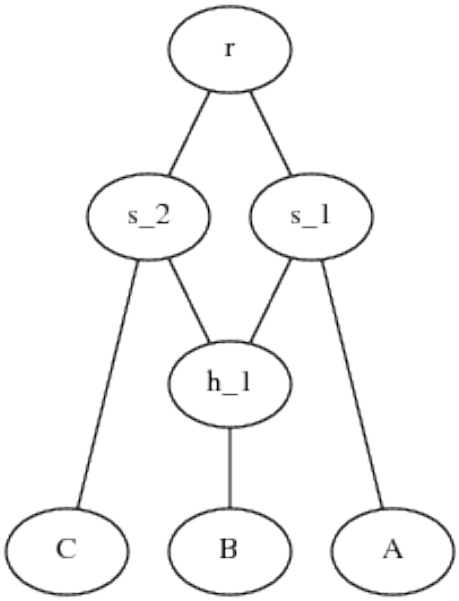
$$((A : t_A, B : t_B) : t_{AB}, C : t_C), \tag{1}$$

where t_i denotes the branch length from i to its parent node in coalescent units.

Extended Newick formatted strings [3, 6] label all internal nodes, and are used for inputting species networks. In the network string, the descendants of a hybrid node are recorded before the hybrid node the first time the hybrid node occurs; otherwise, it is written as a tip node. For example:

$$(((B : t_B)h\#\gamma : t_h^{s1}, A : t_A)s1 : t_{s1}, (h\#\gamma : t_h^{s2}, C : t_C)s2 : t_{s2})r, \tag{2}$$

where # identifies the hybrid node.



At a hybrid node, lineages travel to either parent node with given probabilities. The parameter γ denotes the probability of that the lineage goes to the first parent node. Since the hybrid node has two parent nodes, the branch length needs to be specific, i.e. t_i^j denotes the branch length from i to j in coalescent units.

3.2 Method

The coalescent model of Degnan and Salter [4] is extended to obtain the distribution of gene trees (T) in a given network (W). The network W is initially reduced to a set of simpler networks ($SG(W)$) in a single step of the reduction process. By standard probability theory, we have the following:

$$P(T|W) = \sum_{w^* \in SG(W)} P(T|W^* = w^*)P(W^* = w^*|W)$$

When each network w^* is obtained from W by removing an edge or a node, we will apply the recursion to reduce the networks on W^* until all the simpler network structures are tree-like. The problem of obtaining gene tree probabilities from species trees has already been solved by [4]. The approach we have outlined will therefore reduce the probability of a gene tree, given a species network, into a linear combination of gene tree probabilities, given species trees.

4 Commands:

4.1 Generating a list of all gene tree topologies in a taxa set

```
hybrid-coal -sp INPUT1 -gtopo [-o PREFIX]
```

`INPUT1` is a(n) (extended) Newick formatted string (see Section 3.1), which does not have to be a binary tree. For example, to generate a gene tree topology for the taxon set $\{A, B, C\}$:

```
hybrid-coal -sp '(A,B,C)r;'-gtopo.
```

The default setting will save the gene tree topologies to the file `OUT.topo`. The file name can be specified via the option `-o`, followed by `PREFIX`.

4.2 Calculating gene tree probabilities of a given species network

```
hybrid-coal -sp INPUT1 [-gt INPUT2] [-o PREFIX]
```

`INPUT1` is a(n) (extended) Newick formatted string (see Section 3.1), which can be entered through the command line or a text file.

The flags `-gt` and `INPUT2` are optional. `INPUT2` is a Newick formatted string of a gene tree topology, which can be entered through the command line or from a text file, where users can specify several gene trees. If gene trees are not specified, `hybrid-coal` will generate all possible gene tree topologies and then compute the probabilities. For example:

```
hybrid-coal -sp '((A:1,B:1):1,C:2)r;'
```

will produce the following in `OUT.prob`:

```
1  ((A,C),B)   0.122626
2  (A,(B,C))   0.122626
3  ((A,B),C)   0.754747
```

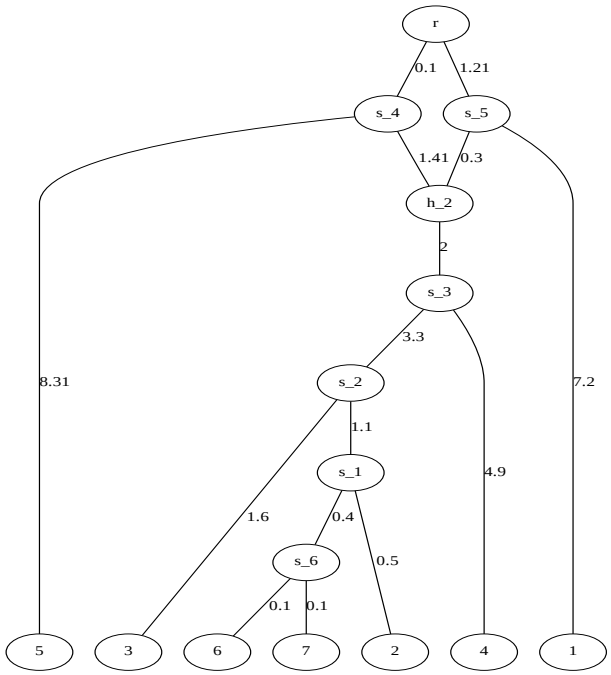
4.3 Commands for other features:

4.3.1 Plot

```
hybrid-coal -sp INPUT -dot OUTPUT [-branch].
```

`hybrid-coal` uses the program `dot` to generate figures. The option `[-branch]` will label the branch lengths in the figure, e.g.:

```
hybrid-coal -sp trees/7_tax_sp_nt1_para -dot -branch.
```



Alternatively, by replacing `-dot` with `plot`, `hybrid-coal` can generate LaTeX code for plotting a network/tree.

5 Summary of command line options

`-h` or `-help`	Help. List the following content.

`-sp INPUT`	Input the species network/tree string through the command line or from a file. Branch lengths of the `INPUT` are in coalescent units.
`-gt INPUT`	Input the gene tree string through the command line or from a file.
`-o STR`	Specify the file name prefix of the output.
`-plot/-dot [option]`	Use LaTeX(`-plot`) or `Dot` (`-dot`) to draw the input (defined by `-sp`) network/tree.

References

[1] Brumfield, R. T. and M. D. Carling (2010). The influence of hybrid zones on species tree inference in manakins. In L. L. Knowles and L. S. Kubatko (Eds.), *Estimating Species Trees, Practical and Theoretical Aspects*, pp. 115–127. Hoboken, NJ: Wiley-Blackwell.

[2] Buckley, T. R., M. Cordeiro, D. C. Marshall, and C. Simon (2006). Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (*Maoricicada Dugdale*). *Systematic Biology 55*(3), 411–425.

[3] Cardona, G., F. Rossell, and G. Valiente (2008). Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics 9*(532-540).

[4] Degnan, J. H. and L. A. Salter (2005). Gene tree distributions under the coalescent process. *Evolution 59*, 24–37.

[5] Holland, B. R., S. Benthin, P. J. Lockhart, V. Moulton, and K. T. Huber (2008). Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evolutionary Biology 8*, 202–213.

[6] Huson, D., R. Rupp, and C. Scornavacca (2010). *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge, UK: Cambridge University Press.

[7] Meng, C. and L. S. Kubatko (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theoretical Population Biology 75*, 35–45.