Hybrid_coal:

Sha (Joe) Zhu

October 5, 2012

1 Introduction

In phylogenetic study, trees are used for describing evolutionary histories. In particular, a species tree presents population divergences, and a gene tree indicates the times that genes start 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 between gene trees and species trees makes describing species relationships very difficult. Common causes of the conflict include gene duplication, horizontal gene transfer, incomplete lineage sorting, and hybridization [5, 7]. If speciation events occur closely, it is likely that some gene copies remain the same after species divergence. This inconsistency between gene trees and species tree is referred as incomplete lineage sorting.

Hybridization refers to interbreeding between species. Offspring who carry gene 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], New Zealand alpine cicadas [2]. Here I propose research on probabilistic modeling of the coalescent with lineage sorting in hybridized species. In these models, relationships between species are represented by a network rather than a tree, while relationships at the gene level are still represented by trees.

`Hybrid_coal` is developted to calculate the gene tree probabilities within a species network.

2 Download and Compilation

`Hybrid_coal` can be downloaded from https://code.google.com/p/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_sim` on UNIX-like systems. In the directory `hybrid_coal-VERSION`, execute the following command:

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

Note: Command `make doxygen-run` will generate an `HTML` documentation of the source code.

3 Notations

3.1 Coalescent parameters

Under the coalescent process, the waiting time of lineages to coalesce is an exponential random variable. The Kingman coalescent process only allows two lineages to coalesce at a time. Thus, the mean of the waiting time that b lineages coalesce to $b-1$ lineages is $\binom{b}{2}$ per unit of time.

3.2 Input/Output formats

The input file for `hybrid_coal` is a character string that describes relationships between species. *Standard Newick* format [8] is used for inputting species trees and outputting gene trees. For example,

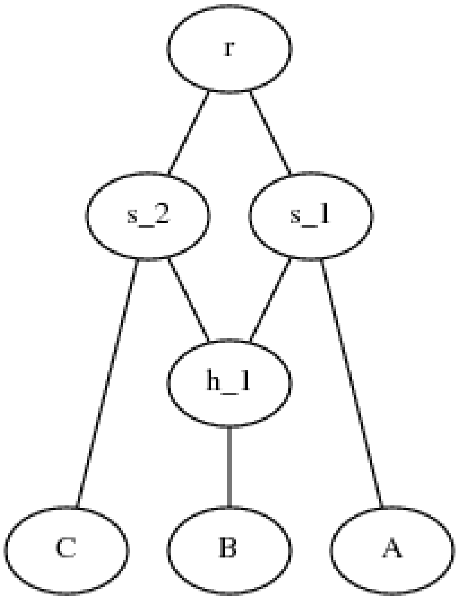
$$((A:t_A, B:t_B):t_{AB}, C:t_C), \tag{1}$$

where t_i denote the the branch length from i to its parent node in coalescent unit.

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, if the hybrid node appears at the first time; 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 # sign identifies the hybrid node.



At a hybrid node, lineages travel to either its parent nodes with some probabilities. 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, that is t_i^j denotes the branch length from i to j in coalescent unit.

3.3 Method

The coalescent model [4] is extended for obtaining the distribution of gene trees (T_g) in a given network (N). Network N is initially reduced to a set of simpler networks ($G(N)$). By standard probability theory, we have the following:

$$P(T_g = \tau|N) = \sum_{N^* \in G(N)} P(T_g = \tau|N^*)\omega(N^*|T_g = \tau, N),$$

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

4 Commands:

4.1 Generate a list of all gene tree topologies of a taxa set

```
hybrid_coal -sp INPUT1 -gtopo

hybrid_coal -sp INPUT1 -gtopoF OUTPUT
```

`INPUT1` is a Newick formated string (see section 3.2), which does not have to be a binary tree. For example, to generate gene tree topology for taxon set $\{A, B, C\}$.

```
hybrid_coal -sp '(A,B,C)r;'-gtopo
```

The default setting will save the gene tree topologies to file `GENE_topo`. The file name can be specified via option `-gtopoF`.

```
(A,C),B);
(A,(B,C));
((A,B),C);
```

To generate gene tree topologies for multiple lineages for the same species, for example, for the same taxon set $\{A, B, C\}$, with 2 lineages of species A, use command:

```
hybrid_coal -sp '(A_1,A_2,B,C)r;' -gtopoF A2BC
```

```
(((A_1,C),B),A_2);
((A_1,(B,C)),A_2);
(((A_1,B),C),A_2);
((A_1,B),(A_2,C));
(((A_1,B),A_2),C);
((A_1,C),(A_2,B));
(A_1,((A_2,C),B));
```

```
(A_1,(A_2,(B,C)));
(A_1,((A_2,B),C));
((A_1,(A_2,B)),C);
(((A_1,C),A_2),B);
((A_1,(A_2,C)),B);
(((A_1,A_2),C),B);
((A_1,A_2),(B,C));
(((A_1,A_2),B),C);
```

4.2 Calculate gene tree probabilities given species network

```
hybrid_coal -sp INPUT1 [-gt INPUT2] [-out OUTPUT]
```

INPUT1 is a(n) (extended) Newick form string (see section 3.2), which can be entered through command line or a text file.

The flag -gt and INPUT2 are optional. INPUT2 is a Newick formated string of a gene tree topology, which can be entered through command line or 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, then compute the probabilities. For example:

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

will print the following message:

```
   1  ((A,C),B)   0.122626
   2  (A,(B,C))   0.122626
   3  ((A,B),C)   0.754747
       Total        1
Species Input: ((A:1,B:1):1,C:2)r;
Species structure: ((A:1,B:1):1,C:2)r;
Total probability: 1
Gene tree probabilities produced in file: out_coal
```

The gene tree probabilities are saved in file out_coal by default. Users can specify the filename via option -out.

4.3 Generate Maple script of the gene tree probabilities

Hybrid_coal can also generate Maple script to calculate the gene tree probabilities. Option -symb enables users to calculate the symbolic probabilities of the gene trees for analytic work. By default, the Maple script is saved in file maple_prob.mw. Users can specify the filename via option -mapleF.

```
hybrid_coal -sp INPUT1 [-gt INPUT2] -maple [-symb]

hybrid_coal -sp INPUT1 [-gt INPUT2] -mapleF OUTPUT
```

4.4 Generate coalescent histories of the gene tree probabilities

Hybrid_coal can also generate an extensive LaTeXcode for user to study the coalescent history of a gene tree within a network.

```
hybrid_coal -sp INPUT1 [-gt INPUT2] -latex

hybrid_coal -sp INPUT1 [-gt INPUT2] -latexF OUTPUT
```

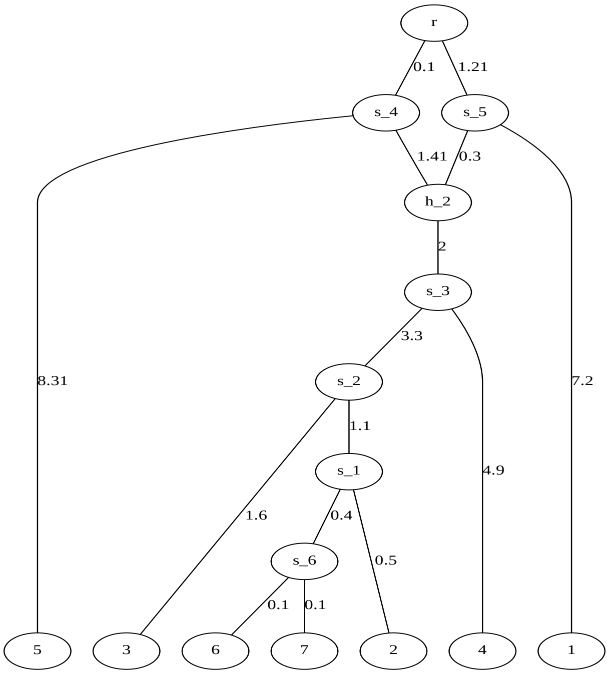
4.5 Commands for other features:

4.5.1 Plot

```
hybrid_coal -sp INPUT -dotF OUTPUT [-branch]
```

Hybrid_coal uses program dot to generate figures. Options [-branch] will label the branch lengths in the figure. For example,

```
hybrid_coal -sp trees/7_tax_sp_nt1_para -dot -branch
```



If option `-dot` is used instead of `-dotF OUTPUT`, figure will be saved in file `figure.pdf` by default.

Alternatively, by replacing `-dotF` with `plotF`, `hybrid_coal` can generate LaTeXcode for plotting a network/tree. If `-plot` is used instead of `-plotF OUTPUT`, LaTeXcode will be saved in file `texfigure.tex` by default.

5 Summary of command line options

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

`-sp INPUT` Input the species network/tree string through command line or a file. Branch lengths of the `INPUT` are in coalescent unit.

`-gt INPUT` Input the gene tree string through command line or a file.

`-latex` / `-latexF` Generate the coalescent history of a gene tree within a species network.

`-maple/ -mapleF` Generate a `Maple` executeable script file to calculate the gene tree probabilities given species networks.

`-symb` To enable the `Maple` script calculate the symbolic gene tree probabilities.

`-gtopo / -gtopoF` To generate the gene tree topologies of a given set of taxa.

`-plot/-dot [option]` Use LaTeX(`-plot`) or `Dot` (`-dot`) to draw the input (defined by `-sp`) network(tree).

`-plotF/-dotF FILE` Generated figure will be saved in `FILE`.

References

[1] Robb T. Brumfield and Matthew. D. Carling. The influence of hybrid zones on species tree inference in manakins. In L. Lacey Knowles and Laura S. Kubatko, editors, *Estimating Species Trees, Practical and Theoretical Aspects*, pages 115–127. Wiley-BlackWell, 2010.

[2] Thomas. R. Buckley, Michael. Cordeiro, David. C. Marshall, and Chris. Simon. Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (*maoricicada dugdale*). *Systematic Biology*, 55(3):411–425, 2006.

[3] Gabriel. Cardona1, Francesc. Rossell, and Gabriel. Valiente. Extended newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics*, 9(532), 2008.

[4] James H. Degnan and Laura A. Salter. Gene tree distributions under the coalescent process. *Evolution*, 59:24–37, 2005.

[5] Barbara R. Holland, Steffi Benthin, Peter J. Lockhart, Vincent Moulton, and Katharina T. Huber. Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evolutionary Biology*, 8:202, 2008.

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

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

[8] Gary Olsen. Gary olsen's interpretation of the "Newick's 8:45" tree format standard, 1990. http://evolution.genetics.washington.edu/phylip/newick_doc.html.

[9] Sha Zhu, James H. Degnan, and Mike Steel. Probabilistic modeling of gene trees given species networks, 2011. Poster, http://www.newton.ac.uk/programmes/PLG/Zhu.pdf.