Overview

Our simulator does the following:

Step 1: Generate a parent tree under the Multispecies Coalescent (MSC).

Step 2: Generate duplication and loss events through a top-down birth-death process on the species tree.

- When a duplication is generated, draw a daughter tree from the MSC, place the duplication event on the daughter tree, update the number of available lineages, and continue the birth-death process in the species tree.
- When a loss is generated, place the loss on any available subtree (the parent tree or any existing daughter tree), update the number of available lineages, and continue the birth-death process in the species tree.

Step 3: Coalesce the parent and daughter trees to generate a gene family tree.

Model Summary

We begin by generating duplication and loss events through a top-down birth-death process on the species tree. Briefly, we draw the timing of events from an exponential distribution with a scale parameter $1/ ((\lambda + \mu)*N)$. Then, we determine whether the event was a duplication or a loss based on λ and μ. If the event is a duplication, we draw a daughter tree from the Multispecies Coalescent (MSC) and place the duplication event on the daughter tree. The duplication event can be placed on any edge of the daughter tree that is present in the correct branch of the species tree at the time of duplication. The daughter tree is truncated above the event, and all unsampled branches are discarded. If the event is a loss, we place the loss on an available tree (the parent tree or any daughter trees with edges present in the correct branch of the species tree), and truncate the tree. We keep track of N (the number of available copies) by tracking the number of subtrees with edges present in each branch of the species tree. After we have generated all subtrees and placed duplications and losses, we coalesce the subtrees. We begin with the parent tree, and then sort trees by age. We coalesce the oldest trees first. Subtrees can coalesce to any branch in the growing parent tree that exists at the relevant time in the relevant branch of the species tree.

Algorithm Description

1. **Input:** Specify a species tree (with branch lengths in coalescent units), μ, and λ.

2. **Birth-death:** Run a top-down birth-death process in the species tree as follows:
   - Generate a parent gene tree under the MSC. Add it to the list of available gene trees.
   - For each branch in the species tree (beginning at the root):
     - Record the number of copies N at the beginning of the branch. At the root, start with one. Set $t_{current}$ to zero.

- While $t_{current}$ < branch.length and N > 0*:*
  - Draw t_{next} (the time to the next event) from an exponential distribution with scale parameter 1 / ((λ + μ)**N*).
  - Set $t_{current} = t_{current} + t_{next}$.
  - If $t_{current}$ < branch.length:
    - Decide whether the event is a duplication [with probability λ/(λ + μ)] or a loss [with probability μ/(λ + μ)].
    - If the event is a duplication, draw a daughter tree from the MSC. At the time of duplication, sample a single edge present in the correct branch of the species tree. Truncate the gene tree above this point, discarding unsampled edges. Add the new subtree to the list of available subtrees.
    - If the event is a loss, place the loss on a gene tree edge in an available subtree in the correct branch of the species tree at the time of loss. Discard all edges below the loss event, and truncate.
    - Adjust *N* for the current branch by adding (duplication) or subtracting (loss).
- At the end of processing a species tree branch, set *N* on any species tree branches directly subtending that branch. Set *N* based on the number of subtrees that exist in that branch of the species tree.

3. Coalesce subtrees

- Sort subtrees by duplication age (oldest to youngest).
- For each duplicated subtree:
  - Set ***age*** to the age of the duplication event.
  - Until coalescence occurs:
    - Find the branch of the species tree corresponding the duplicated subtree.
    - Until we have a time of coalescence
      - Find all available edges. This will include edges from the parent tree and edges from any subtrees that have previously coalesced. Only those edges that exist in the correct branch of the species tree are considered.
      - Draw a time to coalescence **(*tcoal*)** from an exponential distribution with a rate parameter set to the number of potential edges.
      - If this coalescence event fails to occur before any of the potential edges coalesce with each other:
        - Set ***age*** to the time at which the first edges coalesce and continue looking for the coalescence time.
      - Else, we have found our time of coalescence ***tcoal*.**
    - If **tcoal** occurs in the current species tree branch, then we have coalesced, and we do the following:
      - Select an edge to coalesce to (uniform probability over available edges as determined above).
      - Update taxon namespace in the subtree to reflect the copy number.
      - Adjust edge lengths of the subtree and the parent tree.
      - Combine parent tree and subtree to create the new parent tree.
    - If **tcoal** does not occur in the current species tree branch:

        - Set ***age*** to the age of the current branch of the species tree.
- Return the gene family tree which now contains all copies.