

The Structured Coalescent

Ian Roberts

07/06/2021

The Coalescent

- The Coalescent is a model for the ancestral process of a sample of lineages
- A coalescence event occurs when a pair of lineages within the sample encounter a common ancestor backwards in time
- Each pair of lineages coalesces at constant rate 1, resulting in a combined coalescence rate of $\frac{1}{\theta} \binom{n_t}{2}$ where n_t denotes the number of lineages remaining in the sample at time t backwards in time and θ denotes the product of effective population and mutation rate
- The coalescent process will continue until only a single lineage remains, called the most recent common ancestor (MRCA)

Homochronous Leaves

- Homochronous leaves occur when each lineage is sampled at the same time
- **Homochronous.Sim** simulates the ordinary Kingman n -coalescent (**Kingman1982?**) with simultaneously sampled lineages
- **Homochronous.Sim** outputs a phylo object from the ape package (containing edge matrix, vector of edge lengths)
- Simulations are generated by the following steps:
 - Generate event times T_n, T_{n-1}, \dots, T_2 , for $T_k \sim \text{Exp}\left(\frac{1}{\theta} \binom{k}{2}\right)$
 - Select pairs of lineages remaining in the sample to coalesce at each event time

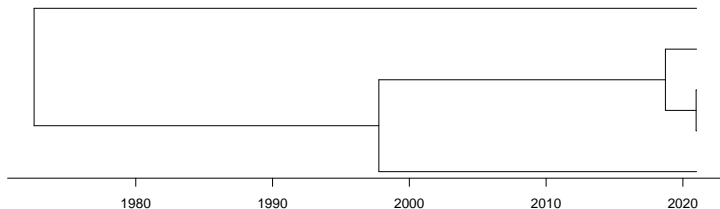


Figure 1: Homochronous n -coalescent simulation using Homochronous.Sim where Effective population \times generation length = 25

Heterochronous Leaves

- Heterochronous leaves occur when lineages are sampled at various times
- **Heterochronous.sim** again simulates the Kingman n -coalescent with lineages sampled at various times. Requires an additional input of the time of the sample measured in continuous years
- Reduces to **Homochronous.sim** if all leaves are sampled at the same time

- Simulations are generated by iterating backwards in time, setting $t = 0$ to be the time of the most recently sampled lineage:
 - Let T denote the time until a new sampled lineage is added to the tree ($T = \infty$ if all lineages are already included)
 - With probability e^{-T} , a new sampled lineage is added before the next coalescence event. Set $\tilde{T} = T$
 - Otherwise, with probability $1 - e^{-T}$, a coalescence event occurs after time \tilde{T} , drawn from an exponential distribution truncated above at T
 - Update $t \leftarrow t + \tilde{T}$
 - Repeat until there is a single lineage remaining and all sampled lineages have been included in the tree

ID	1	2	3
Date	05/01/2021	30/05/2020	14/03/2021
Cts. Time	2021.014	2020.411	2021.200
ID	4	5	
Date	08/09/2018	16/12/2019	
Cts. Time	2018.687	2019.958	

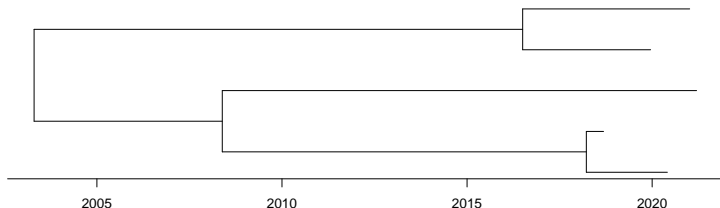


Figure 2: Heterochronous n -coalescent simulation using Homochronous.Sim where Effective population \times generation length = 25

The Structured Coalescent

- Under the structured coalescent, multiple populations or demes evolve with migration events occurring occasionally between distinct demes
- Each pair of lineages coalesce at rate 1, with coalescences occurring at rate $\frac{1}{\theta_i} \binom{n_{it}}{2}$ in deme i , where n_{it} denotes the number of lineages in deme i at time t and θ_i denotes the product of the effective population size of deme i and the mutation rate. Further, each lineage may migrate from deme i to deme j ($j \neq i$) at rate λ_{ij}
- **Structured.sim** generates a realisation of this structured coalescent process with heterochronous leaves, where each event is one of:
 - Coalescence: Two lineages in the same deme merge
 - Migration: One lineage moves between two demes

- Simulations are generated by iterating backwards in time, setting $t = 0$ to be the time of the most recently sampled lineage:
 - Let T denote the time until a new sampled lineage is added to the tree ($T = \infty$ if all lineages are already included)
 - With probability e^{-T} , a new sampled lineage is added before the next coalescence or migration event. Set $\tilde{T} = T$
 - Otherwise, with probability $1 - e^{-T}$, either a coalescence or migration event occurs after time \tilde{T} , drawn from an exponential distribution truncated above at T . The type of event is selected based on the relative rates of coalescence and migration events
 - Update $t \leftarrow t + \tilde{T}$
 - Repeat until there is a single lineage remaining and all sampled lineages have been included in the tree

Example Simulation

- Augment the previous example dataset with an initial deme for each lineage
- Assume 2 demes and an example migration matrix Λ

ID	Age	Deme
1	2021.014	1
2	2020.411	1
3	2021.200	2
4	2018.687	2
5	2019.958	2

$$\Lambda = \begin{bmatrix} 0 & 0.02 \\ 0.05 & 0 \end{bmatrix}$$

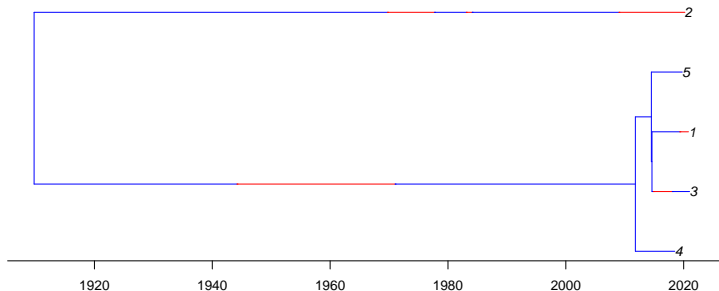


Figure 3: Structured heterochronous n -coalescent realisation using Structured.sim with Effective population \times generation length = 25 and migration rates 0.02 ($1 \rightarrow 2$) and 0.05 ($2 \rightarrow 1$)

Heterochronous.Sim

- Simulations obtained from **Heterochronous.sim** were verified via calculation of the log-likelihood
- The log-likelihood can be calculated recursively by considering the contributions made when either a new leaf is added, or a coalescence occurs
- This log-likelihood can be compared to the likelihood from Drummond et al. (2002),

$$L = \frac{1}{\theta^{n-1}} \prod_{i=2}^{2n-1} \exp \left\{ -\frac{k_i(k_i - 1)}{2\theta} (t_i - t_{i-1}) \right\}$$

where

- t_i denotes the time of the i^{th} event (either a coalescence or adding a newly sampled lineage)
- k_i denotes the number of current lineages at t_i
- θ denotes the product of effective population and generation length.

Structured.Sim

- Simulations from **Structured.Sim** were verified via calculation of the log-likelihood with the exact form of the likelihood from Ewing, Nicholls, and Rodrigo (2004),

$$L = \prod_{i=1}^d \prod_{j \neq i} \prod_{r=2}^{2n+M-1} \frac{1}{\theta_i^{c_i}} \lambda_{ij}^{m_{ij}} \exp \left\{ - \left(\frac{k_{ir}(k_{ir} - 1)}{2\theta_i} + k_{ir} \lambda_{ij} \right) (t_r - t_{r-1}) \right\}$$

- where
 - n denotes the number of leaves
 - d denotes the number of demes
 - m gives a $d \times d$ matrix with entries m_{ij} giving the number of migrations ($i \rightarrow j$) and $M = \sum_{i,j=1}^d m_{ij}$ gives the total number of migration events
 - c gives a d -dimensional vector with entries c_i giving the number of coalescence events occurring in deme i
 - k_{ir} denotes the number of lineages in deme i at time t_r

Recursive log-likelihood

The recursive log-likelihood for the structured coalescent can be calculated by summing the contributions when each node is added to the phylo. Nodes can be added in one of three ways:

- A new leaf is added
- A coalescence occurs in deme i
- A migration occurs $i \rightarrow j$

Likelihood contributions

Total coalescence rate $\mathcal{C} = \sum_{i=1}^d \frac{1}{\theta_i} \binom{k_i}{2}$, Total migration rate $\mathcal{M} = \sum_{i=1}^d \sum_{j \neq i} k_i \lambda_{ij}$ and Total event rate: $\mathcal{E} = \mathcal{C} + \mathcal{M}$.

- New leaf added:
 - $L \leftarrow L + \log(1 - \mathcal{E} \exp\{-\mathcal{E}\delta\})$ where δ is the time to next new leaf being added
- Coalescence event in deme i :

$$\begin{aligned} L &\leftarrow L + \log(\mathbb{P}[\text{Coal}]\mathbb{P}[\text{Coal deme}]\mathbb{P}[\text{Coal pair}]f_T(t)) \\ &= L + \log\left(\frac{\mathcal{C}}{\mathcal{E}} \cdot \frac{\frac{1}{\theta_i} \binom{k_i}{2}}{\sum_{j=1}^d \frac{1}{\theta_j} \binom{k_j}{2}} \cdot \binom{k_i}{2}^{-1} \cdot \mathcal{E} \exp\{-\mathcal{E}t\}\right) = L - \mathcal{E}t - \log \theta_i \end{aligned}$$

- Migration event $i \rightarrow j$:

$$\begin{aligned} L &\leftarrow L + \log(\mathbb{P}[\text{Mig}]\mathbb{P}[\text{Mig origin}]\mathbb{P}[\text{Mig target}]\mathbb{P}[\text{Mig lineage}]f_T(t)) \\ &= L + \log\left(\frac{\mathcal{M}}{\mathcal{E}} \cdot \frac{k_i \sum_{\eta \neq i} \lambda_{i\eta}}{\sum_{\alpha=1}^d k_{\alpha} \sum_{\beta \neq \alpha} \lambda_{\alpha\beta}} \cdot \frac{\lambda_{ij}}{\sum_{\gamma \neq i} \lambda_{i\gamma}} \cdot \frac{1}{k_i} \cdot \mathcal{E} \exp\{-\mathcal{E}t\}\right) = L - \mathcal{E}t + \log(\lambda_{ij}) \end{aligned}$$

Testing

Running 100 generated processes for the example dataset and comparing the exact likelihood (Ewing, Nicholls, and Rodrigo 2004) to the recursive likelihood. Maximum discrepancy over the 100 iterations was $7.1054274 \times 10^{-15}$.

```
N <- 100; likelihoods <- matrix(0,nrow = N,ncol = 2)

for (i in 1:N){
  phylo <- Structured.sim(struc.data, 1,1,2,migration.mat,F
  likelihoods[i,1] <- phylo$log.likelihood
  likelihoods[i,2] <- structured.likelihood(phylo,1,1,migra
}

max(abs(likelihoods[,2] - likelihoods[,1]))

## [1] 7.105427e-15
```


Aim

Currently implementing a reversible-jump MCMC scheme to infer the migration events and migration rates for a fixed genealogy.
Based on the work of:

- Drummond et al. (2002) who propose a reversible-jump MCMC scheme to simultaneously estimate the genealogy alongside the mutation rate and effective population size for a heterochronous coalescent model
- Ewing, Nicholls, and Rodrigo (2004) who develop this MCMC scheme to again simultaneously estimate the genealogy alongside the mutation rate, population size and migration rates for a structured coalescent model

Reversible Jump Moves

The selected reversible jump moves are those of Ewing, Nicholls, and Rodrigo (2004) which do not modify the (fixed) genealogy:

- Migration Birth/Death
- Migration Pair Birth/Death
- Coalescent Node Merge/Split

Migration Birth/Death

A migration birth/death move samples an ancestral node, r , uniformly at random and selects either a birth or death event with probability $\frac{1}{2}$,

- Birth event: Add one new migration node uniformly along the edge connecting r and $\text{parent}(r)$, reselecting demes over the maximal connected subtree containing the edge $\langle r, \text{parent}(r) \rangle$ containing no interior migration nodes
- Death event: If the $\text{parent}(r)$ is a migration node, delete $\text{parent}(r)$ and reselect demes over the maximal connected subtree containing the edge $\langle r, \text{parent}^2(r) \rangle$ containing no interior migration nodes

Migration Pair Birth/Death

A migration pair birth/death move samples an edge, e , uniformly at random and selects either a birth or death event with probability $\frac{1}{2}$,

- Pair birth event: A pair of migration nodes are each added uniformly along e , selecting a deme different to the deme of e
- Pair death event: If both terminal nodes of e are migration nodes and lie in the same deme, both nodes are deleted and the deme between them is updated

Coalescent Node Merge/Split

A coalescent node merge/split samples a coalescent node, c , uniformly at random and selects either a merge or split event with probability $\frac{1}{2}$,

- Merge event: If both children of c are migration nodes lying in the same deme, the two migration nodes are pulled through c and merged onto the parent edge of c
- Split event: If $\text{parent}(c)$ is a migration node, the migration node is pulled through c and split onto both child edges

References

- Drummond, Alexei J, Geoff K Nicholls, Allen G Rodrigo, and Wiremu Solomon. 2002. "Estimating Mutation Parameters, Population History and Genealogy Simultaneously from Temporally Spaced Sequence Data." *Genetics* 161 (3): 1307–20.
- Ewing, Greg, Geoff Nicholls, and Allen Rodrigo. 2004. "Using Temporally Spaced Sequences to Simultaneously Estimate Migration Rates, Mutation Rate and Population Sizes in Measurably Evolving Populations." *Genetics* 168 (4): 2407–20.