

The Structured Coalescent

Ian Roberts

07/06/2021

Homochronous Leaves

Homochronous.Sim simulates the ordinary Kingman n -coalescent, and outputs a phylo object which can potentially be plotted using `ape::plot.phylo`

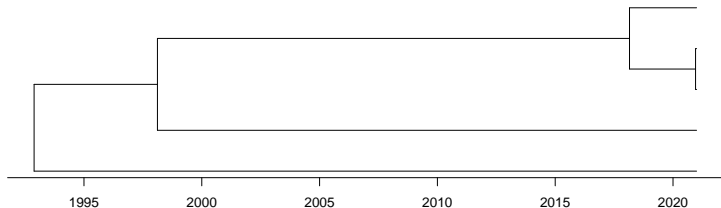


Figure 1: Homochronous n -coalescent simulation using **Homochronous.Sim** where Effective population \times generation length = 25

Heterochronous Leaves

- **Heterochronous.sim** again simulates the Kingman n -coalescent, outputting a phylo object. Key difference is the option for heterochronous leaves (in continuous years) and optional argument for plotting the phylo object automatically
- Effectively reduces to **Homochronous.sim** if all leaves have the same age
- Example dataset: first column giving sample id and second column giving year of observation

ID	Year
1	2021
2	2020
3	2021
4	2018
5	2019

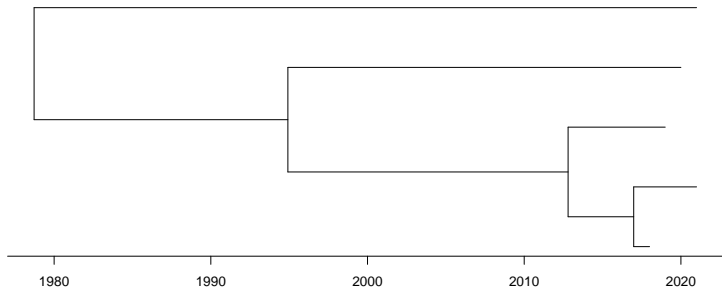


Figure 2: Heterochronous n -coalescent simulation using Homochronous.Sim where Effective population \times generation length = 25

Verification

- Verified the accuracy of **Heterochronous.sim** via likelihood computation. Likelihood calculated recursively as the simulation is run, then compare obtained value with explicit calculation from the final phylo object (using **phylo.likelihood**).
- Likelihood of a heterochronous coalescent process (Drummond et al. 2002) is given by

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

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 Coalescent

- **Structured.sim** generates a realisation of the structured coalescent process where at each event time, either:
 - Two lineages in the same deme coalesce
 - One lineage migrates between two demes
- Allows for heterochronous leaves so effectively reduces to **Heterochronous.sim** when there is only one deme
- Example dataset: Augments the previous dataset with a third column giving initial deme of the lineages. Assume 2 demes and an example migration matrix Λ

ID	Year	Deme
1	2021	1
2	2020	1
3	2021	2
4	2018	2
5	2019	2

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

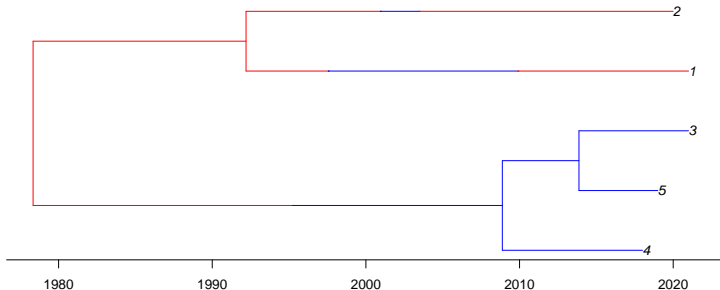


Figure 3: Structured heterochronous n -coalescent realisation using Structured.sim with Effective population \times generation length = 25 and migration rates 0.02 ($1 \rightarrow 2$) and 0.05 ($2 \rightarrow 1$)

Verification (IN PROGRESS)

- Again, verify by constructing likelihood recursively and compare to the exact form (Ewing, Nicholls, and Rodrigo 2004)

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

- 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

Problem

Currently likelihood calculated recursively does not agree with the exact form from Ewing, Nicholls, and Rodrigo (2004). Recursive likelihood is calculated by adding a contribution from each of the three possible events:

- New leaf is added
- Coalescence occurs in deme i
- Migration occurs $i \rightarrow j$

Likelihood contributions

Total coalescence rate $\mathcal{C} = \sum_{i=1}^d \binom{k_i}{2}$, 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 or migration event:
 - Event time T is simulated as truncated exponential with rate \mathcal{E} and cutoff δ via inverse transform sampling
 - With probability $\frac{\mathcal{C}}{\mathcal{M}}$, event is a coalescence. Select a coalescence deme from all demes containing at least 2 lineages weighted by number of lineages in the deme so deme i is selected with probability $\frac{k_i}{\sum_{j: k_j \geq 2} k_j}$.

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.