HOMEWORK 2: COALESCENT WITH MUTATION

VLADIMIR SHCHUR

Deadline: November 14th. Google doc (to submit your homework) `https://docs.google.com/document/d/1mf82FFpyebBf5kyTULM1d6Dl7TqgjL1VrRNkyqz1dz0/edit?usp=sharing`.

1. Basics

Chromosome is an interval $[0, 1]$.

Individual (or individual's genome) is a set of M chromosomes, numbered from 0 to $M - 1$.

Chromosomes with the same *id* (from different individuals) are related by a single tree genealogy (no recombination).

Genealogies for chromosomes with different *id*s are simulated independently from each other.

2. Coalescence with mutation

Let there be K lineages. Mutation rate is μ, effective population size over time is $\nu(t)$. Assume that $\nu(t)$ is piecewise constant function.

Coalescence with mutation is a Poisson process with the (variable) rate

$$\omega(K, t) = K\mu + \frac{1}{\nu(t)}\binom{K}{2}.$$

Simulation scheme.

(1) Set $t = 0$, initialise K.
(2) Sample time T till the next event from Poisson process with the rate $\omega(K, t)$. Set $t = t + T$.
(3) Generate type of the event following Bernoulli distribution with weights proportional to $K\mu$ (mutation) and $\frac{1}{\nu(t)}\binom{K}{2}$ (coalescence).
   - Mutation: sample ancestral lineages ℓ where mutation occurs independently from K available lineages. Sample mutation position p on a genome uniformly on $[0, 1]$. All individuals which are decedents of ℓ get variant 1 at position p. All other individuals have variant 0 at position p.
   - Coalescence. Choose uniformly a random pair of lineages ℓ_1 and ℓ_2. These two lineages coalesce at time t. Update genealogy. Set $K = K - 1$.
(4) stop if $K = 1$. Otherwise go to step 2.

3. Debugging/verification

Hudson's ms simulator can be used to verify results. `http://home.uchicago.edu/~rhudson1/source/mksamples.html`.

Possible simulation scenarios for $K = 2, M = 100000$. `./ms 2 100000 -t 1 -T -eN 0 3 -eN 0.025 0.1 -eN 0.325 1.5 -eN 3 3` (population sizes are $3, 0.1, 1.5, 3$, change at times $0.05, 0.65, 6$) and `./ms 2 100000 -t 1 -T -eN 0 1.5 -eN 3 3` (population sizes are $1.5, 3$, change at time 6). **NB**[1] times in ms are twice as small as in your simulator (but doublecheck it!).

The first version of your simulator can be coalescence without mutation ($\mu = 0$), and you output only coalescent times. Take 2 individuals, hence you simulate only one coalescent event per a pair of chromosomes. Start with constant effective population size $\nu(t) = 1$ and ms simulation `./ms 2 100000 -t 0 -T`. Compare the distribution of coalescent times from your simulator and ms. Then try the same scenario with 3 individuals `./ms 3 100000 -t 0 -T`.

Now add mutation ($\mu > 0$). Take $K > 2$ (e.g. 10 or even 100). Adjust the mutation rate, so that the number of mutation on each genealogy is big (e.g. 10^5). Check that on each edge of a genealogy the number of mutations is approximately proportional to the edge's length (the number of mutations on the edge is a Poisson variable with the rate proportional to the edges length).

4. Comments

Use `seed()` (or similar) to enable results replication.

[1] NB - nota bene (latin).