

# Time Machine Algorithm

## 1 Notation

In the following, the mutation rate is denoted by  $\mu$ , and the  $d \times d$  transition matrix between types by  $\mathbf{P} = (p_{ij})$ . Moreover, for convenience of notation, we denote by  $\mathbf{e}_i$  the  $d$ -dimensional row vector with 1 in the  $i^{\text{th}}$  position and 0 elsewhere. Finally, for a  $d$ -dimensional vector  $\mathbf{x}$ , we write  $|\mathbf{x}|_1 = |x_1| + \dots + |x_d|$ .

## 2 Algorithm

The algorithm simulates genealogical trees backwards in time from an observed initial population  $\mathbf{n}_0 = (n_{1,0}, \dots, n_{d,0})$  with  $d$  possible types up to the point when there are  $N$  sequences left in the population. Note that the case  $N = 1$  corresponds to the ordinary coalescent simulation, whereas  $N > 1$  corresponds to the Time Machine.

Starting at  $t = 0$ , the following steps will be iterated until  $|\mathbf{n}_t|_1 = N$ .

1. Sample the offspring type  $i$  with probability proportional to  $n_{i,t}$ ;
2. Sample the ancestor type  $j$ ; an offspring of type  $i$  might have arisen from an ancestor of type  $j$  through:
  - (a) a coalescent event, with probability proportional to  $|\mathbf{n}_{i,t}|_1 - 1$ ;
  - (b) a  $j \rightarrow i$  mutation (with  $j$  possibly equal to  $i$ ), with probability  $\mu \kappa_{ij} p_{ij}$ , where

$$\kappa_{ij} = \begin{cases} \frac{n_{j,t} + \mu \pi_j}{|\mathbf{n}_t|_1 - 1 + \mu} & j \neq i, \\ \frac{n_{j,t} - 1 + \mu \pi_j}{|\mathbf{n}_t|_1 - 1 + \mu} & j = i, \end{cases}$$

and  $\pi$  is the stationary distribution associated with  $\mathbf{P}$ ;

3. Update the population size within each type,

$$\mathbf{n}_{t+1} = \begin{cases} \mathbf{n}_t - \mathbf{e}_i + \mathbf{e}_j & \text{if a mutation occurred,} \\ \mathbf{n}_t - \mathbf{e}_i & \text{if a coalescent event occurred;} \end{cases}$$

4. Compute the contribution to the likelihood of the simulated event, which is given by

$$w_t = \frac{K_t}{K_{t+1}} \frac{\kappa_{ii} x_{j,t+1}}{\kappa_{ij} |\mathbf{n}_t|_1}$$

where  $K_t = |\mathbf{n}_t|_1 (|\mathbf{n}_t|_1 - 1 + \mu)$ , if a mutation occurred, and by

$$w_t = \frac{K_t}{K_{t+1}} \frac{1}{\kappa_{ii}} \frac{x_{i,t+1} (|\mathbf{n}_{t+1}|_1 - 1)}{n_{i,t} (n_{i,t} - 1)}$$

if a coalescent event occurred;

5. Update the log-likelihood,

$$W_t = \begin{cases} \log(w_t) & t = 0, \\ W_{t-1} + \log(w_t) & t \geq 1; \end{cases}$$

6. Assess the stopping criterion:

- (a) if the Time Machine is used ( $N > 1$ ), stop if  $|\mathbf{n}_{t+1}|_1 = N$ ;
- (b) otherwise, repeat the above steps until  $|\mathbf{n}_{t+1}|_1 = 2$ , at which point mutations are simulated until both remaining sequences are of the same type.

For  $N > 1$ , the log-likelihood is corrected by adding the following term,

$$\log b = \log \left[ \frac{|\mathbf{n}_\rho|_1! \Gamma(\mu)}{\Gamma(\mu + |\mathbf{n}_\rho|_1)} \right] + \sum_{i=1}^d \log \left[ \frac{\Gamma(n_{i,\rho} + \mu \pi_i)}{n_{i,\rho}! \Gamma(\mu \pi_i)} \right],$$

where  $\rho$  is the last simulated event, and  $\Gamma$  denotes the gamma function.