Continuous-Time Stochastic Model for Microsatellite Evolution

1. Overview

For each microsatellite locus, we model the change in length over time as a Continuous-Time Markov Chain (CTMC) on the integer state space $L_t \in \mathbb{Z}$, where $L_t = 0$ denotes the ancestral length.

2. CTMC Definition

Each microsatellite locus evolves independently according to:

$$L \to \begin{cases} L+1 & \text{at rate } \mu_i, \\ L-1 & \text{at rate } \mu_d. \end{cases}$$

where:

μ_i : insertion rate (1-bp addition), μ_d : deletion rate (1-bp loss), L_t : net length change (bp), t : continuous time.

2.1 Generator Matrix

The infinitesimal generator $Q = (q_{ij})$ is:

$$q_{ij} = \begin{cases} \mu_i, & j = i+1, \\ \mu_d, & j = i-1, \\ -(\mu_i + \mu_d), & j = i, \\ 0, & \text{otherwise.} \end{cases}$$

The system evolves by the Kolmogorov forward equation:

$$\frac{dP(L,t)}{dt} = \mu_i P(L-1,t) + \mu_d P(L+1,t) - (\mu_i + \mu_d)P(L,t),$$

with initial condition $P(L,0) = \delta_{L,0}$.

2.2 Analytical Behavior

Expected mean and variance evolution:

$$\mathbb{E}[L(t)] = (\mu_i - \mu_d)t, \qquad \text{Var}[L(t)] = (\mu_i + \mu_d)t.$$

Variance grows linearly with time, characteristic of diffusion-like stochastic motion. In cancer evolution, this process is non-equilibrium and demonstrates cumulative drift.

3. Gillespie Simulation Algorithm

The Gillespie algorithm enables exact stochastic simulation of the CTMC. For N loci evolving independently:

3.1 Event Dynamics

Event rate per locus:

$$\lambda_j = \mu_i + \mu_d.$$

Total rate:

$$\Lambda = \sum_{j=1}^{N} \lambda_j = N(\mu_i + \mu_d).$$

Next-event time:

$$\tau \sim \text{Exp}(\Lambda),$$

with mean waiting time $1/\Lambda$. Random locus selection:

$$j^* \sim \text{Uniform}\{1, \ldots, N\}.$$

Event type:

$$\begin{cases} L_{j^*} \to L_{j^*} + 1 & \text{with prob. } \mu_i/(\mu_i + \mu_d), \\ L_{j^*} \to L_{j^*} - 1 & \text{with prob. } \mu_d/(\mu_i + \mu_d). \end{cases}$$

Update:

$$t \leftarrow t + \tau.$$

Repeat until $t \geq T_{\text{total}}$. This generates statistically exact CTMC trajectories of microsatellite evolution.

3.2 Expected Behavior Across Loci

Each locus follows identical independent dynamics, giving population distribution:

$$P_N(L, t) = \frac{1}{N} \sum_{j=1}^{N} \mathbb{P}(L_t^j = L).$$

This mean approximates the analytical distribution $P(L, t)$.

5. Approximate Bayesian Computation (ABC)

The aim is to infer posterior parameter distributions for

$$\theta = (\mu_i, \mu_d, T_{\text{total}}),$$

from observed microsatellite lengths L_{obs}. Since direct likelihood evaluation is intractable, an ABC framework is applied.

5.1 Priors

$$\mu_i \sim \mathcal{U}(\mu_{i,\min}, \mu_{i,\max}), \quad \mu_d \sim \mathcal{U}(\mu_{d,\min}, \mu_{d,\max}), \quad T_{\text{total}} \sim \mathcal{U}(T_{\min}, T_{\max}).$$

5.2 Distance Metric

Similarity measured by 1-Wasserstein distance:

$$D(L_{\text{sim}}, L_{\text{obs}}) = \int_{-\infty}^{\infty} |F_{\text{sim}}(x) - F_{\text{obs}}(x)| \, dx.$$

Smaller D indicates closer agreement.

5.3 Sequential Monte Carlo (SMC) Algorithm

At generation $g = 1, \ldots, G$:

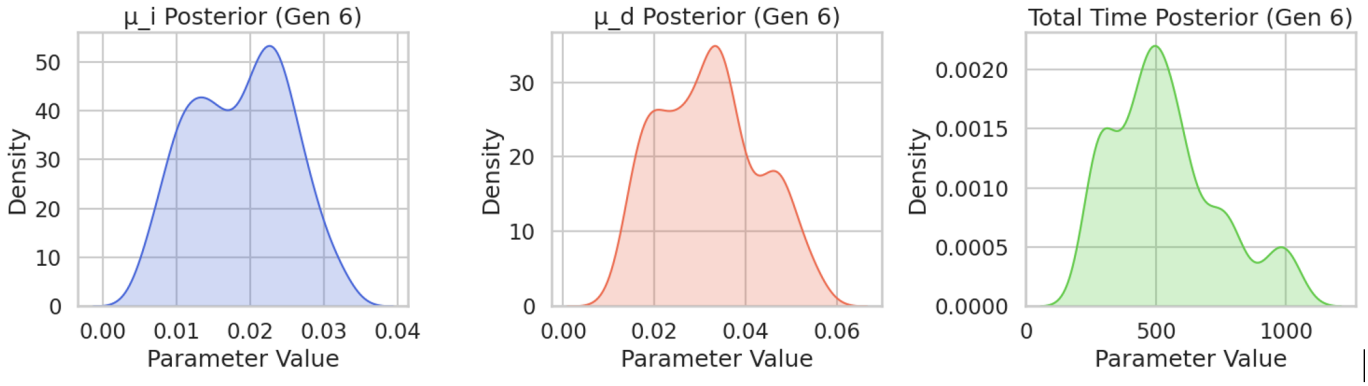
1. Sample N_p particles $\theta_i^{(g)}$.
2. Simulate data $L_{\text{sim}}^{(i)} \sim f(\theta_i^{(g)})$.
3. Compute $D_i = D(L_{\text{sim}}^{(i)}, L_{\text{obs}})$.
4. Retain $D_i < \varepsilon_g$.
5. Update threshold: $\varepsilon_{g+1} = \text{median}(D_i) \times \alpha,\ \alpha < 1$.
6. Weighting: $w_i \propto e^{-D_i/\varepsilon_g},\ \sum_i w_i = 1$.
7. Resample and jitter parameters:
$$\theta_i^{(g+1)} \leftarrow \theta_i^{(g)} + \mathcal{N}(0, \sigma_{\text{jit}}^2).$$

Final weighted particles approximate the ABC posterior:

$$p(\theta|L_{\text{obs}}) \approx \text{ABC posterior}.$$

Results

```
Generation 6: 100%|██████████| 400/400 [00:23<00:00, 17.38it/s]
```




```
Parameter Estimates:
True values: μ_i = 0.015, μ_d = 0.025, Total time = 600
Estimated:  μ_i = 0.0186, μ_d = 0.0318, Total time = 533.45
```

Figure 1: Posterior Parameter Distributions of the Final Inference Iterationx