Mathematical Description of the Microsatellite Mutation Framework

1. Overview

Equilibrium Model

Assumes that the distribution of microsatellite lengths is at stationary equilibrium, as typically assumed in germline population genetics studies.

Dynamic Model

Describes microsatellite evolution as a non-equilibrium, stochastic Markov process over a finite number of effective cell divisions D_{eff}, characteristic of an expanding tumor population.

Both models parameterize indel dynamics through insertion and deletion rates μ_i and μ_d, which drive length transitions in loci over cell divisions.

2. Equilibrium Model

2.1 Birth–Death Representation

Under the equilibrium assumption, the microsatellite length $L \in \{-L_{\max}, \dots, L_{\max}\}$ evolves through a birth–death Markov process:

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

The rate matrix for this process is

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

The equilibrium distribution $\pi(L)$ satisfies

$$\pi Q = 0, \quad \sum_L \pi(L) = 1.$$

This yields a geometric stationary distribution with ratio parameter $\rho = \mu_i/\mu_d$:

$$\pi(L) \propto \rho^L, \quad \text{for } |L| \leq L_{\max}.$$

The detailed balance condition,

$$\mu_i \, \pi(L) = \mu_d \, \pi(L+1),$$

ensures no net drift in length.

2.2 Sampling from the Stationary Distribution

At equilibrium, the distribution of microsatellite lengths across N loci is given by multinomial sampling from $\pi(L)$:

$$P(\mathbf{L}) = \prod_{j=1}^{N} \pi(L_j),$$

where $\mathbf{L} = \{L_1, L_2, \ldots, L_N\}$. Simulation of equilibrium data proceeds by sampling $L_j \sim \pi(L)$, approximating the steady-state distribution observed in population microsatellites.

3. Dynamic Microsatellite Model (Non-Equilibrium)

3.1 Model Description

Tumor cells undergo clonal expansion with ongoing mutations. Microsatellite length distribution evolves over a finite number of effective divisions D_{eff}, representing total mitotic events since the tumor's origin. The per-locus state L_t^j evolves in continuous time via stochastic 1-bp insertions and deletions. Transition probabilities per infinitesimal time interval Δt are

$$\begin{cases} P(L \to L+1) = \mu_i \, \Delta t, \\ P(L \to L-1) = \mu_d \, \Delta t, \\ P(L \to L) = 1 - (\mu_i + \mu_d)\Delta t. \end{cases}$$

The time evolution follows the master 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 $P(L,0) = \delta_{L,0}$.

3.2 Discrete-Time Approximation

The model discretizes time with step size

$$\Delta t = \min\{0.1, 1/\max(\mu_i, \mu_d)\},$$

and number of steps $N_{\text{steps}} = D_{\text{eff}}/\Delta t$. At each step:

$$L_{t+\Delta t}^j = L_t^j + X_t^j,$$

where X_t^j is drawn from

$$X_t^j = \begin{cases} +1, & \text{with prob. } \mu_i \Delta t, \\ -1, & \text{with prob. } \mu_d \Delta t, \\ 0, & \text{with prob. } 1 - (\mu_i + \mu_d)\Delta t. \end{cases}$$

The final microsatellite distribution after D_{eff} divisions is $\{L^j(D_{\text{eff}})\}_{j=1}^{N}$.

3.3 Non-Equilibrium Property

Expectations over time evolve as:

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

The non-equilibrium process continues to evolve and does not reach stationarity under exponential tumor growth.

4. Inference Framework: ABC–SMC

4.1 Priors

$$\mu_i \sim \mathcal{U}(\mu_{i,\min}, \mu_{i,\max}), \quad \mu_d \sim \mathcal{U}(\mu_{d,\min}, \mu_{d,\max}), \quad D_{\text{eff}} \sim \mathcal{U}(D_{\min}, D_{\max}) \text{ (dynamic model only).}$$

4.2 Distance Metric

The discrepancy between simulated and observed data is quantified by the 1-Wasserstein distance:

$$W(P_{\text{sim}}, P_{\text{obs}}) = \int_{-\infty}^{\infty} |F_{\text{sim}}(x) - F_{\text{obs}}(x)|\, dx,$$

where F_{sim} and F_{obs} are the cumulative distributions.

4.3 SMC Sampling

At iteration k, draw N_p parameter samples $\{\theta_i^{(k)}\} = (\mu_i, \mu_d, D_{\text{eff}})$. For each:

$$L_{\text{sim}}^{(i)} = f_{\text{model}}(\theta_i^{(k)}), \quad d_i = W(L_{\text{obs}}, L_{\text{sim}}^{(i)}).$$

Accept particles with $d_i < \epsilon_k$, where

$$\epsilon_{k+1} = \text{Quantile}_{50\%}(d_i^{(k)}) \times \text{decay factor}.$$

Resample accepted particles with weights $w_i \propto \exp(-d_i/\epsilon_k)$, adding Gaussian jitter to maintain diversity. The final weighted ensemble approximates the posterior:

$$p(\mu_i, \mu_d, D_{\text{eff}} \mid L_{\text{obs}}) \propto \text{ABC posterior}.$$

5. Equilibrium vs. Dynamic Model Comparison

Posterior parameter means under each assumption:

$$(\hat{\mu}_i^{\text{eq}}, \hat{\mu}_d^{\text{eq}}) \quad \text{and} \quad (\hat{\mu}_i^{\text{dyn}}, \hat{\mu}_d^{\text{dyn}}, \hat{D}_{\text{eff}}^{\text{dyn}}).$$

Relative bias:

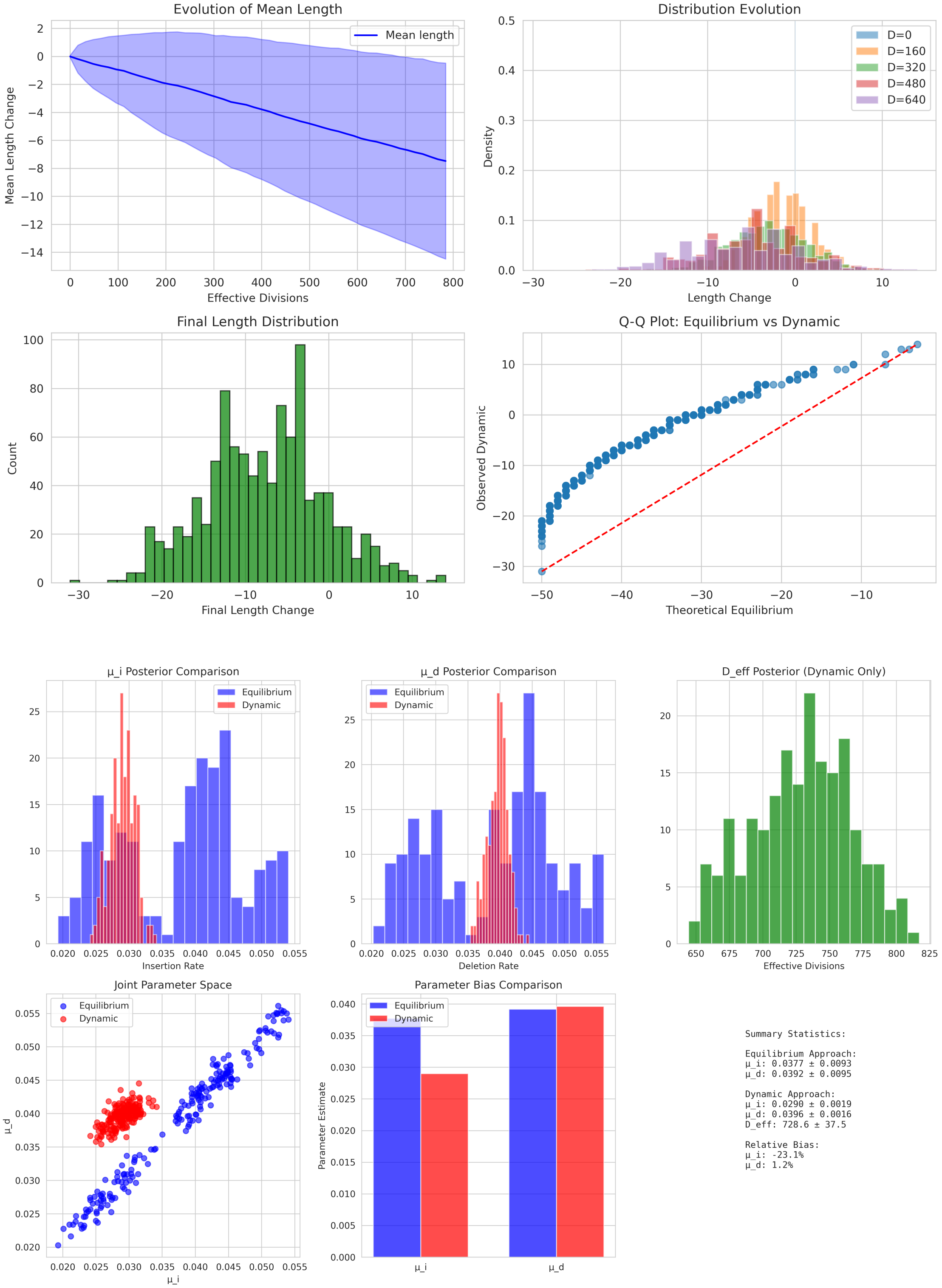
$$\text{Bias}_\mu = \frac{\hat{\mu}^{\text{dyn}} - \hat{\mu}^{\text{eq}}}{\hat{\mu}^{\text{eq}}}.$$

Comparing these posteriors measures the bias from assuming equilibrium dynamics.
In the limit $D_{\text{eff}} \to \infty$,

$$\lim_{t\to\infty} P(L, t; \mu_i, \mu_d) = \pi(L).$$

The equilibrium microsatellite model is a steady-state limit of the dynamic stochastic process. Testing it quantifies bias in equilibrium assumptions when applied to tumor populations in non-equilibrium scenario.

Results



```
Summary Statistics:

Equilibrium Approach:
μ_i: 0.0377 ± 0.0093
μ_d: 0.0392 ± 0.0095

Dynamic Approach:
μ_i: 0.0290 ± 0.0019
μ_d: 0.0396 ± 0.0016
D_eff: 728.6 ± 37.5

Relative Bias:
μ_i: -23.1%
μ_d: 1.2%
```

Figure 1: True parameters: $\mu_i = 0.025$, $\mu_d = 0.035$, $D_{\text{eff}} = 800$