Importance Sampling for Context-Dependent Evolutionary Models

by

Kewei Xu

Department of Statistical Science

Duke University

Defense Date: July 15th, 2025

Approved:

Scott C. Schmidler, Supervisor

Kevin J. Wiehe

Alexander Fisher

A thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2025

ABSTRACT

Importance Sampling for Context-Dependent Evolutionary Models

by

Kewei Xu

Department of Statistical Science

Duke University

Defense Date: July 15th, 2025

Approved:

Scott C. Schmidler, Supervisor

Kevin J. Wiehe

Alexander Fisher

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistical Science in the Graduate School of Duke University

2025

Abstract

This thesis focuses on optimizing and applying the importance sampling algorithm for context-dependent evolutionary models. First, we use variational inference to update the parameters of the independent-site model as proposal distribution to optimize the importance sampling algorithm. Then, we try blockwise importance sampling algorithm to optimize the importance sampling algorithm. Finally we apply the importance sampling to estimate unknown parameters given the start and end sequences over a specified time interval.

Contents

List of Tables

List of Figures

Acknowledgement

First of all, I want to thank my thesis advisor, Scott C. Schmidler, for his constructive suggestions during this research project. His guidance has always inspired me to finish this thesis.

I want to thank Kevin J. Wiehe and Alexander Fisher for serving on my committee. I would also like to send my appreciation to Joseph Mathews for his feedback on my progress and Yongkang Li for sharing his ideas.

1 Introduction

Evolutionary models are fundamental for understanding genetic variation and mutation patterns over time. DNA sequence evolution is most commonly modeled as a continuous-time Markov chain process (CTMC) using independent-site evolutionary models that assume each nucleotide evolves independently [5, 6]. Context-dependent evolutionary models, which account for such dependencies, present computational challenges, especially when estimating transition probabilities $P(\mathbf{y}|\mathbf{x})$ over time T. We could tackle this kind of problem using importance sampling.

This thesis aims to extend previous work [1,2,8] by optimizing an importance sampling method for context-dependent evolutionary models and utilizing it for parameter estimation. Specifically, in Chapter 4 and 5, we try to optimize the importance sampling to estimate $P(\mathbf{y}|\mathbf{x})$ for context-dependent models using variational inference and blockwise importance sampling. In Chapter 6, we use importance sampling to estimate unknown parameters—given the starting and ending sequences of the evolutionary process—via the IS-EM algorithm.

2 Review

DNA sequence evolution is a traditional problem, particularly the calculation of transition probability that sequence $\mathbf{x} = (x_1, \ldots, x_n)$ evolves into sequence $\mathbf{y} = (y_1, \ldots, y_n)$ over the time T. A simplistic model treats each base as evolving independently by a continuous-time Markov chain (CTMC) with a rate matrix Q [5, 6]. Specifically, the transition probability would be [2]:

$$P(\mathbf{y} \mid \mathbf{x}) = \prod_{i=1}^{n} (e^{TQ})_{x_i, y_i}$$

However, in reality, site-specific mutations don't follow the independent-site assumption. In particular, in immune system processes (like somatic hypermutation), the likelihood of a mutation at a particular site is strongly affected by its neighboring bases. Certain short DNA motifs are known to trigger higher mutation frequencies because they are better targeted by enzymes [8, 11].

To account for dependencies between neighboring sites, we extend the current model so that site-specific mutation rates depend on the context window centered on each site [10]. With $k \geq 1$, we define the context window surrounding the center position i as [2]:

$$\tilde{x}_i = (x_{i-k}, ..., x_i, \ldots, x_{i+k}),$$

We define $\gamma(b; \tilde{x}_i)$ as substitution rate from base x_i to a new base b given the context window [2]. The cumulative mutation rate for the entire sequence $\mathbf{x}$ would be [2]:

$$\Gamma(\mathbf{x}) = \sum_{i=1}^{n} \sum_{b \neq x_i} \gamma(b; \tilde{x}_i)$$

We model the entire evolutionary process over the time interval $[0, T]$ from sequence $\mathbf{x}$ to sequence $\mathbf{y}$. Each evolutionary path involves a number of discrete mutations. Specifically, an evolutionary path is denoted by [2]:

$$P = \{(t_1, l_1, b_1), (t_2, l_2, b_2), ..., (t_{m-1}, l_{m-1}, b_{m-1}), (t_m, l_m, b_m)\}$$

The t_j represents the time point of the jth mutation, the l_j represents the location of

jth mutation, and the b_j represents the new base for jth mutation. Let $\mathbf{x}^{(j)}$ indicate the intermediate sequence after jth mutations. The expression of the path likelihood is [2]:

$$\mathcal{L}(P) = \left\{ \prod_{j=1}^{m} \left[\gamma(b_j; \tilde{x}_{l_j}^{(j-1)}) \cdot e^{-\Gamma(\mathbf{x}^{(j-1)})(t_j - t_{j-1})} \right] \right\} \cdot e^{-\Gamma(\mathbf{x^{(m)}})(T - t_m)} 1_{\mathbf{x^{(m)}}=\mathbf{y}}.$$

However, direct computation of the transition probability for context-dependent model is difficult because it is necessary to integrate over all possible paths. In Chapter 3, we introduce an importance sampling technique to tackle this issue.

3 Importance Sampling

Computing the transition probability from sequence $\mathbf{x}$ to sequence $\mathbf{y}$ under a context-dependent model is computationally difficult. To solve this problem, we introduce an importance sampling technique which draws mutation paths from a tractable model and uses importance weights to correct for the difference between the context-dependent model and the proposal model [2, 8].

3.1 Transition Probability

Let $\mathcal{P}$ denote the collection of all mutation paths that sequence $\mathbf{x}$ evolves into sequence $\mathbf{y}$ over a time interval $[0, T]$. Each path represents a sequence of mutation events, including their time of mutations, positions of mutations, and nucleotide substitutions. For each path $P \in \mathcal{P}$, we define its likelihood under the context-dependent model as $\mathcal{L}(P)$. Then, the transition probability for context-dependent model would be:

$$\tilde{P}(\mathbf{y} \mid \mathbf{x}) = \int_{P \in \mathcal{P}} \mathcal{L}(P)\, dP$$

Since the integral above cannot be computed directly, we start considering a tractable, site-independent mutation model that does not involve context-dependence. Let $q(P)$ denote the likelihood of path P under this simpler model, which can be simulated directly. Then:

$$\tilde{P}(\mathbf{y} \mid \mathbf{x}) = \int_{P \in \mathcal{P}} \frac{\mathcal{L}(P)}{q(P)} q(P)\, dP$$

Let $P_Q(\mathbf{y}|\mathbf{x}) = \int_{P \in \mathcal{P}} q(P)\, dP$ denote the transition probability for the independent model. Then, the normalized probability distribution of an evolutionary path would be:

$$q(P|\mathbf{y}, \mathbf{x}) = \frac{q(P)}{P_Q(\mathbf{y}|\mathbf{x})}$$

The transition probability for the context-dependent model would be:

$$\tilde{P}(\mathbf{y} \mid \mathbf{x}) = P_Q(\mathbf{y}|\mathbf{x}) \cdot \mathbb{E}_{q(P|\mathbf{y},\mathbf{x})} \left[\frac{\mathcal{L}(P)}{q(P)} \right]$$

We draw independent paths $P^{(1)}, P^{(2)}, \ldots, P^{(N)} \sim q(P|\mathbf{y}, \mathbf{x})$. The unbiased estimator of $\tilde{P}(\mathbf{y} \mid \mathbf{x})$ would be:

$$\widehat{\tilde{P}(\mathbf{y} \mid \mathbf{x})} = \frac{P_Q(\mathbf{y}|\mathbf{x})}{N} \sum_{i=1}^{N} \mathcal{W}(P^{(i)})$$

Summary of Key Notations:

- $\mathcal{L}(P^{(i)})$: Likelihood of a path $P^{(i)}$ under the context-dependent model.
- $q(P^{(i)})$: Likelihood of a path $P^{(i)}$ under the independent model.
- $\mathcal{W}(P^{(i)}) = \frac{\mathcal{L}(P^{(i)})}{q(P^{(i)})}$: The importance weight.

3.2 Effective Sample Size

We use the Effective Sample Size (ESS) to quantifies how many independent samples from the target distribution would provide the same variance of the estimator as the current set of weighted samples. With N weighted samples with normalized importance weights $w_1, w_2, \ldots, w_N$ (with $w_i = \frac{\mathcal{W}(P^{(i)})}{\sum_{j=1}^{N} \mathcal{W}(P^{(j)})}$), the ESS is:

$$\text{ESS} = \frac{1}{\sum_{i=1}^{N} w_i^2}.$$

A higher ESS indicates that the weighted samples are more uniformly distributed, which means that the estimator is more reliable.

3.3 Optimize the Importance Sampling

While we could use importance sampling to approximate the context-dependent transition probability, its efficiency is highly sensitive to the variance of importance weights Large variance implies that only a small number of samples contribute meaningfully to the estimate, resulting in a low effective sample size (ESS).

In this thesis, we are going to improve the performance of the importance sampling procedure by enhancing weight stability and maximizing ESS with the following 2 strategies:

- **Variational Inference:** We optimize parameters in the rate matrix Q (used in independent model) by maximizing the evidence lower bound (ELBO). This allows

the probability distribution of paths sampled from the proposal distribution to better approximate the probability distribution of paths sampled from the context-dependent model, thereby reducing the variance of the estimates.

- **Blockwise Sampling:** We split the whole sequence into nonoverlapping blocks and sample evolutionary paths in a blockwise manner. Blockwise proposals better capture dependencies among neighboring sites, thereby reducing estimator variance.

4 Variational Inference

In this section, we try variational inference to optimize the rate matrix Q for the independent model, thereby reducing the variance of the importance weights and increasing the effective sample size (ESS).

Variational inference aims to construct a tractable distribution that approximates the true distribution for the context-dependent model [3,4]. In this section, we use the Jukes-Cantor model (JC69) model [12] as our independent proposal distribution because it is tractable. The JC69 model assumes that all nucleotide substitutions are equally likely. The rate matrix for the independent JC69 model is:

$$Q = \begin{bmatrix} -3\mu & \mu & \mu & \mu \\ \mu & -3\mu & \mu & \mu \\ \mu & \mu & -3\mu & \mu \\ \mu & \mu & \mu & -3\mu \end{bmatrix}$$

We begin with an independent-site mutation model, where the base substitution rate is parameterized by μ. The context-dependent model we are using in this section is CpG model. To capture context-dependencies in CpG model, the parameter ϕ is introduced to modulate the rate μ when it generates any CG dinucleotides. For CpG model, the mutation rate from the base x_i to a different base $b \neq x_i$ at position i is defined as:

$$\gamma(b; \tilde{x}_i) = \begin{cases} \mu\phi, & \text{if the mutation from } x_i \to b \text{ generate a CG dinucleotide.} \\ \mu, & \text{otherwise.} \end{cases}$$

Our objective is to optimize all parameters—specifically μ (the rate of base substitution) — so that the distribution it induces over mutation paths closely approximates the true posterior distribution defined by the context-dependent CpG mutation process. We could do it by minimizing the KL-divergence between the proposal distribution and the posterior distribution [3, 4].

4.1 KL-Divergence and Evidence Lower Bound

Assume the time interval [0, T], where T is fixed, we have an intractable probability density function of mutation path P given sequences $\mathbf{x}$ and $\mathbf{y}$ for context-dependent model [2]:

$$\tilde{P}(P|\mathbf{y},\mathbf{x}) = \frac{\tilde{P}(\mathbf{y},P|\mathbf{x})}{\int_{P\in\mathcal{P}}\tilde{P}(\mathbf{y},P|\mathbf{x})dP} = \frac{\mathcal{L}(P)}{\tilde{P}(\mathbf{y}\mid\mathbf{x})}$$

Which

$$\mathcal{L}(P) = \tilde{P}(\mathbf{y},P|\mathbf{x})$$

$$\tilde{P}(\mathbf{y}\mid\mathbf{x}) = \int_{P\in\mathcal{P}}\tilde{P}(\mathbf{y},P|\mathbf{x})dP = \int_{P\in\mathcal{P}}\mathcal{L}(P)dP$$

To approximate posterior path distribution $\tilde{P}(P|\mathbf{y},\mathbf{x})$, we use independent-site proposal $q_\mu(P|\mathbf{y},\mathbf{x})$ to approximate $\tilde{P}(P|\mathbf{y},\mathbf{x})$ by finding the optimized value of μ, with

$$q_\mu(P|\mathbf{y},\mathbf{x}) = \frac{q_\mu(\mathbf{y},P|\mathbf{x})}{P_Q(\mathbf{y}|\mathbf{x})}$$

Which

$$P_Q(\mathbf{y}|\mathbf{x}) = \int_{P\in\mathcal{P}} q_\mu(\mathbf{y},P|\mathbf{x})dP$$

As an independent-site proposal, $P_Q(\mathbf{y}|\mathbf{x})$ has a closed form and could be computed directly.

Then, under the measure space of end-point CTMC with initial sequence $\mathbf{x}$ to end sequence $\mathbf{y}$ with fixed time interval [0, T]:

$$\log\tilde{P}(\mathbf{y}|\mathbf{x}) = \log\int_{P\in\mathcal{P}}\tilde{P}(\mathbf{y},P|\mathbf{x})dP$$

$$= \log\int_{P\in\mathcal{P}}\tilde{P}(\mathbf{y},P|\mathbf{x})\frac{q_\mu(P|\mathbf{y},\mathbf{x})}{q_\mu(P|\mathbf{y},\mathbf{x})}dP$$

$$= \log\mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})}\left[\frac{\tilde{P}(\mathbf{y},P|\mathbf{x})}{q_\mu(P|\mathbf{y},\mathbf{x})}\right]$$

$$\geq \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})}\left[\log\left(\frac{\tilde{P}(\mathbf{y},P|\mathbf{x})}{q_\mu(P|\mathbf{y},\mathbf{x})}\right)\right]$$

$$\geq \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})}\left[\log \tilde{P}(\mathbf{y}, P|\mathbf{x})\right] - \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})}\left[\log q_\mu(P|\mathbf{y},\mathbf{x})\right]$$

$$= F(q,\mu)$$

Denote $F(q,\mu)$ as Evidence Lower Bound (ELBO). And the KL-Divergence of path distribution would be:

$$\mathrm{KL}\left[q_\mu(P|\mathbf{y},\mathbf{x})||\tilde{P}(P|\mathbf{y},\mathbf{x})\right] = \int_{P\in\mathcal{P}} q_\mu(P|\mathbf{y},\mathbf{x}) \log\left(\frac{q_\mu(P|\mathbf{y},\mathbf{x})}{\tilde{P}(P|\mathbf{y},\mathbf{x})}\right) dP$$

$$= \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})}\left[\log\left(\frac{q_\mu(P|\mathbf{y},\mathbf{x})}{\tilde{P}(P|\mathbf{y},\mathbf{x})}\right)\right]$$

$$= \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})}\left[\log\left(q_\mu(P|\mathbf{y},\mathbf{x})\right)\right] - \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})}\left[\log\left(\tilde{P}(P|\mathbf{y},\mathbf{x})\right)\right]$$

$$= \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})}\left[\log\left(q_\mu(P|\mathbf{y},\mathbf{x})\right)\right] - \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})}\left[\log\left(\frac{\tilde{P}(\mathbf{y},P|\mathbf{x})}{\tilde{P}(\mathbf{y}|\mathbf{x})}\right)\right]$$

$$= \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})}\left[\log\left(q_\mu(P|\mathbf{y},\mathbf{x})\right)\right] - \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})}\left[\log\left(\tilde{P}(\mathbf{y},P|\mathbf{x})\right)\right] + \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})}\left[\log\left(\tilde{P}(\mathbf{y}|\mathbf{x})\right)\right]$$

$$= -F(q,\mu) + \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})}\left[\log\left(\tilde{P}(\mathbf{y}|\mathbf{x})\right)\right]$$

Since $\log\left[\tilde{P}(\mathbf{y}|\mathbf{x})\right]$ is fixed, we could minimize $\mathrm{KL}\left[q_\mu(P|\mathbf{y},\mathbf{x})||\tilde{P}(P|\mathbf{y},\mathbf{x})\right]$ by maximizing the ELBO term, $F(q,\mu)$. To optimize the variational parameter μ, we maximize the Evidence Lower Bound (ELBO), which could be estimated using Monte Carlo approximation (with expectation term):

$$F(q,\mu) = \int q_\mu(P|\mathbf{y},\mathbf{x};\mu)\log\frac{\tilde{P}(\mathbf{y},P|\mathbf{x})}{q_\mu(P|\mathbf{y},\mathbf{x};\mu)}dP$$

$$= \int q_\mu(P|\mathbf{y},\mathbf{x};\mu)\log\tilde{P}(\mathbf{y},P|\mathbf{x};\mu)dP - \int q_\mu(P|\mathbf{y},\mathbf{x};\mu)\log q_\mu(P|\mathbf{y},\mathbf{x};\mu)dP$$

$$= \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})}\left[\log\tilde{P}(\mathbf{y},P|\mathbf{x}) - \log q_\mu(P|\mathbf{y},\mathbf{x};\mu)\right]$$

$$= \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})}\left[\log\mathcal{L}(P) - \log q_\mu(P|\mathbf{y},\mathbf{x};\mu)\right]$$

$$= \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})}\left[\log\mathcal{L}(P) - (\log q_\mu(\mathbf{y},P|\mathbf{x};\mu) - \log P_Q(\mathbf{y}|\mathbf{x}))\right]$$

For CpG model, with original parameters μ_{ori} and ϕ_{ori} (fixed), let $1_{\mathrm{CG}}(l_j, b_j; \mathbf{x}^{(j-1)})$

indicate the mutation at location l_j to base b_j creates a CG dinucleotide in the sequence $\mathbf{x}^{(j-1)}$, the likelihood function (and log-likelihood) of path P would be:

$$\mathcal{L}(P) = \left\{ \prod_{j=1}^{m} \left[\mu_{ori} \phi_{ori}^{1_{\mathrm{CG}}(l_j, b_j; \mathbf{x}^{(j-1)})} \cdot e^{-(t_j - t_{j-1}) \sum_{i=1}^{n} \sum_{b \neq x_i^{(j-1)}} \mu_{ori} \phi_{ori}^{1_{\mathrm{CG}}(i,b;\mathbf{x}^{(j-1)})}} \right] \right\} \times e^{-(T-t_m) \sum_{i=1}^{n} \sum_{b \neq x_i^{(m)}} \mu_{ori} \phi_{ori}^{1_{\mathrm{CG}}(i,b;\mathbf{x}^{(\mathbf{m})})}} 1_{\mathbf{x}^{(\mathbf{m})}=\mathbf{y}}.$$

$$\log \mathcal{L}(P) = \left[\log \left(\mu_{ori}^{m} \, \phi_{ori}^{\sum_{j=1}^{m} 1_{\mathrm{CG}}(l_j, b_j; \mathbf{x}^{(j-1)})} \right) - \mu_{ori} \sum_{j=1}^{m} (t_j - t_{j-1}) \sum_{i=1}^{n} \sum_{b \neq x_i^{(j-1)}} \phi_{ori}^{1_{\mathrm{CG}}(i,b;\mathbf{x}^{(j-1)})} - \mu_{ori}(T - t_m) \sum_{i=1}^{n} \sum_{b \neq x_i^{(m)}} \phi_{ori}^{1_{\mathrm{CG}}(i,b;\mathbf{x}^{(\mathbf{m})})} \right] 1_{\mathbf{x}^{(\mathbf{m})}=\mathbf{y}}.$$

Under the measure space of end-point CTMC with initial sequence $\mathbf{x}$ to end sequence $\mathbf{y}$ after time T, we assume $1_{\mathbf{x}^{(\mathbf{m})}=\mathbf{y}} = 1$. For JC69 model, with parameters μ which is going to be optimized, the likelihood function (and log-likelihood) of path P would be:

$$q_\mu(\mathbf{y}, P|\mathbf{x}, \mu) = \mu^m \cdot e^{-3n\mu T}.$$

$$\log q_\mu(\mathbf{y}, P|\mathbf{x}, \mu) = \left[m \log(\mu) - 3n\mu T \right].$$

Note that $\tilde{P}(P|\mathbf{y}, \mathbf{x})$ and $q_\mu(P|\mathbf{y}, \mathbf{x})$ doesn't share the same μ parameters. With parameters μ_{opt} and ϕ_{opt} fixed in $\tilde{P}(P|\mathbf{y}, \mathbf{x})$, we update the μ parameter for $q_\mu(P|\mathbf{y}, \mathbf{x})$ and make it approximate the path distribution of $\tilde{P}(P|\mathbf{y}, \mathbf{x})$.

We know that for JC69 model as independent proposal:

- **Non-mutated nucleotides**:

$$P_Q[i, i] = \frac{1}{4} + \frac{3}{4} e^{-4\mu T}$$

- **Mutated nucleotides** ($i \neq j$):

$$P_Q[i,j] = \frac{1}{4} - \frac{1}{4}e^{-4\mu T}$$

Therefore, the ELBO would be:

$$\begin{aligned} F(q,\mu) = \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})} \Bigg\{ \Bigg[& \log(\mu_{ori}^m \phi^{\sum_{j=1}^m 1_{\mathrm{CG}}(l_j,b_j;\mathbf{x}^{(j-1)})}) \\ & - \mu_{ori} \sum_{j=1}^m (t_j - t_{j-1}) \sum_{i=1}^n \sum_{b \neq x_i^{(j-1)}} \phi^{1_{\mathrm{CG}}(i,b;\mathbf{x}^{(j-1)})} \\ & - \mu_{ori}(T - t_m) \sum_{i=1}^n \sum_{b \neq y_i} \phi^{1_{\mathrm{CG}}(i,b;\mathbf{y})} \Bigg] \\ & - \Big[m \log(\mu) - 3n\mu T \Big] + \log\left[P_Q(y|x)\right] \Bigg\} \end{aligned}$$

$$\begin{aligned} = \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})} \Bigg\{ & \Bigg[m \log(\mu_{ori}) + \log\left(\phi^{\sum_{j=1}^m 1_{\mathrm{CG}}(l_j,b_j;\mathbf{x}^{(j-1)})}\right) \\ & -\mu_{ori} \left(\sum_{j=1}^m (t_j - t_{j-1}) \sum_{i=1}^n \sum_{b \neq x_i^{(j-1)}} \phi^{1_{\mathrm{CG}}(i,b;\mathbf{x}^{(j-1)})} + (T - t_m) \sum_{i=1}^n \sum_{b \neq y_i} \phi^{1_{\mathrm{CG}}(i,b;\mathbf{y})} \right) \Bigg] \\ & - [m \log(\mu) - 3n\mu T] \\ & + \log \left[\left(\frac{1}{4} + \frac{3}{4}e^{-4\mu T} \right)^{\sum_{i=1}^n I(y_i = x_i)} \left(\frac{1}{4} - \frac{1}{4}e^{-4\mu T} \right)^{\sum_{i=1}^n I(y_i \neq x_i)} \right] \Bigg\} \end{aligned}$$

$$\begin{aligned} = \mathbb{E}_{q_\mu(P|\mathbf{y},\mathbf{x})} \Bigg\{ & \Bigg[m \log(\mu_{ori}) + \log\left(\phi^{\sum_{j=1}^m 1_{\mathrm{CG}}(l_j,b_j;\mathbf{x}^{(j-1)})}\right) \\ & -\mu_{ori} \left(\sum_{j=1}^m (t_j - t_{j-1}) \sum_{i=1}^n \sum_{b \neq x_i^{(j-1)}} \phi^{1_{\mathrm{CG}}(i,b;\mathbf{x}^{(j-1)})} + (T - t_m) \sum_{i=1}^n \sum_{b \neq y_i} \phi^{1_{\mathrm{CG}}(i,b;\mathbf{y})} \right) \Bigg] \\ & - [m \log(\mu) - 3n\mu T] \\ & + \sum_{i=1}^n I(y_i = x_i) \log\left(\frac{1}{4} + \frac{3}{4}e^{-4\mu T} \right) + \sum_{i=1}^n I(y_i \neq x_i) \log\left(\frac{1}{4} - \frac{1}{4}e^{-4\mu T} \right) \Bigg\} \end{aligned}$$

With endpoint-conditioned CTMC sampling [7], we could draw $P^{(1)}, P^{(2)}, \cdots, P^{(N)} \sim$

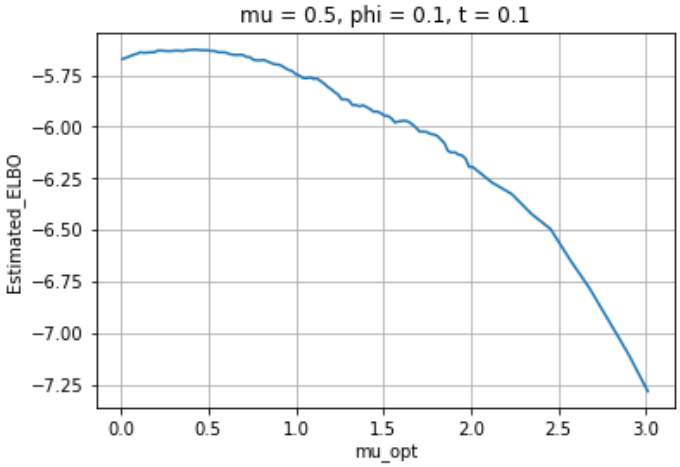
$q_\mu(P|\mathbf{y}, \mathbf{x})$. From these samples, we could get the Monte Carlo estimate of the evidence lower bound (ELBO) for a specified μ and find the optimal μ which would maximize the ELBO. In the next section, we would apply it to a simulation study.

4.2 Simulation Study

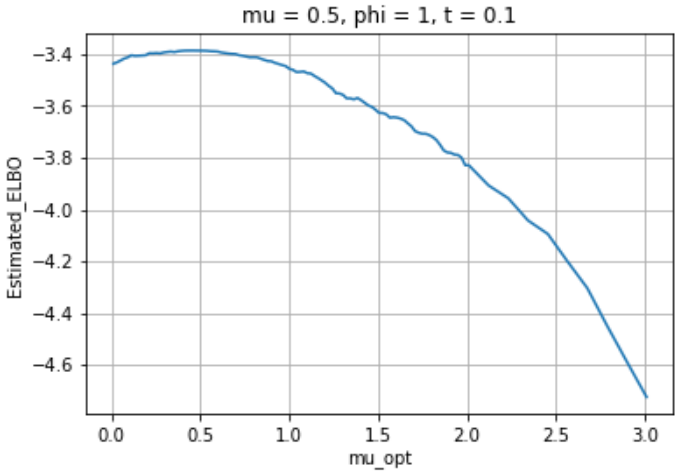
4.2.1 Short Sequence (3-bp)

For our simulation study, we try a simple example with a 3-base pair (3-bp) sequence transition with the starting sequence $\mathbf{x} = \text{CCA}$ and the ending sequence $\mathbf{y} = \text{CCT}$. For the CpG model, we fix the mutation rate at $\mu_{ori} = 0.5$, set the end time to $T = 0.1$, and use different levels of the context-dependence parameter $\phi_{ori} = [0.1, 1, 10]$ to simulate a large number of evolutionary paths for CpG model. We plot the estimated ELBO with respect to μ from the JC69 model as the proposal distribution. Figure 1 shows the relationship between estimated ELBO and μ for different values of the context-dependence parameter ϕ_{ori}.

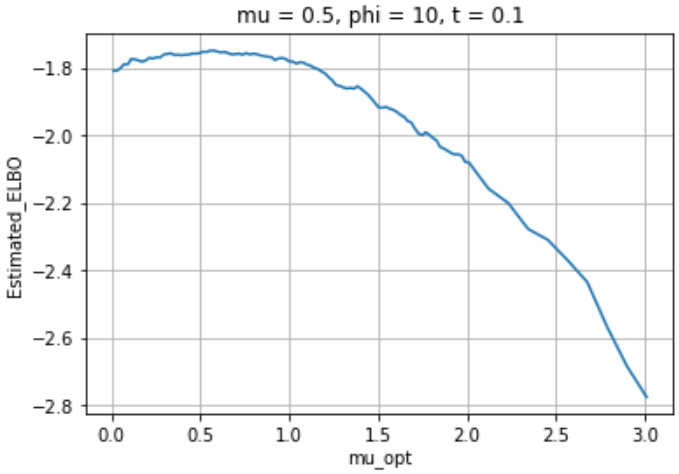
Next, we estimate the transition probabilities $\tilde{P}(\mathbf{y} \mid \mathbf{x})$ multiple times under the CpG model using importance sampling, with different values of the parameter μ used in the JC69 model as the proposal. For each value of $\mu \in [0, 2]$, we do 10 trials and compare these estimates to the exact transition probability. We then compare the variability of these estimates to assess how the parameter μ affects the estimation accuracy. Figure 2 shows the relationship between the estimated transition probabilities and μ under different values of ϕ_{ori}. The red horizontal line indicates the exact transition probability computed by full $4^3 \times 4^3$ instantaneous rate matrix.



(a) $\phi_{ori} = 0.1$

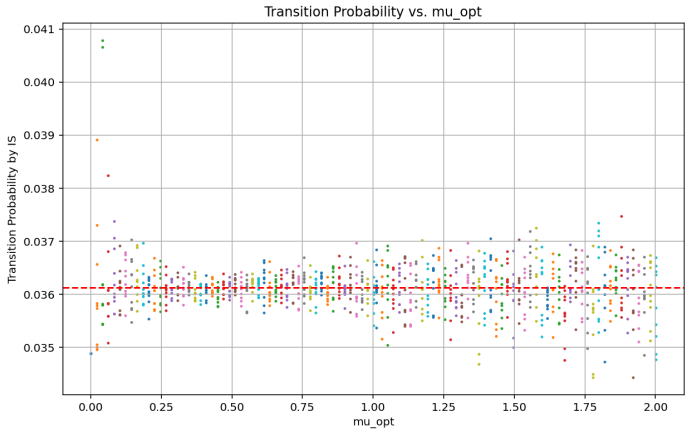


(b) $\phi_{ori} = 1$

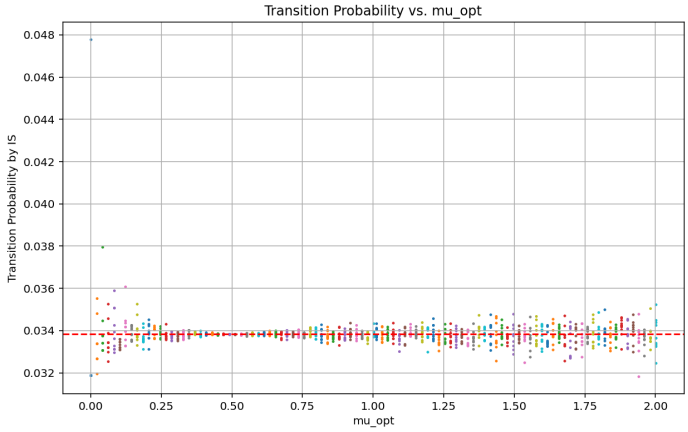


(c) $\phi_{ori} = 10$

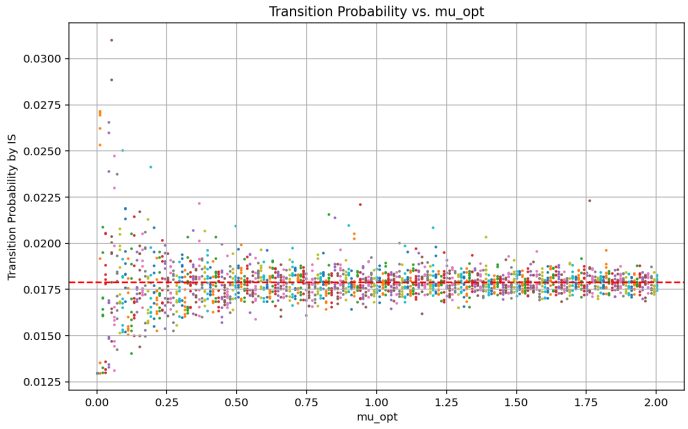
Figure 1: Estimated ELBO vs. μ under different values of ϕ_{ori}.



(a) $\phi_{ori} = 0.1$



(b) $\phi_{ori} = 1$



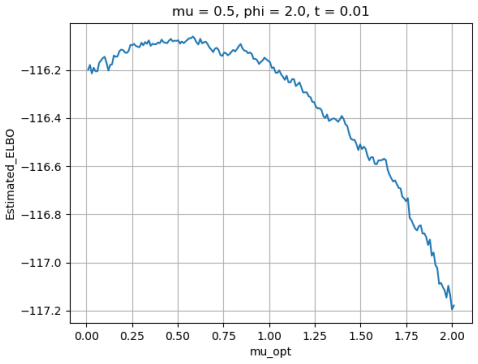
(c) $\phi_{ori} = 10$

Figure 2: Estimated Transition Probabilities vs. μ under different values of ϕ_{ori}.

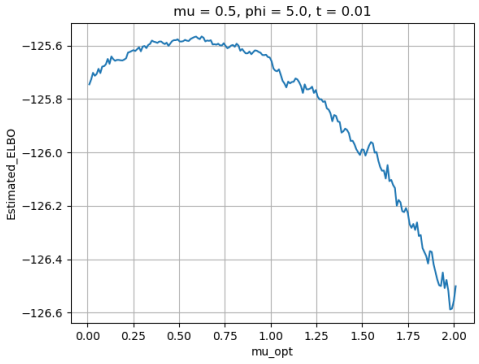
From figures 1 and 2, we could see that when the value of μ in proposal maximizes the estimated ELBO, the variance of our estimates of transition probabilities in importance sampling could be reduced to some extent. It demonstrates that the optimal parameter μ_{opt} in proposal that maximizes the ELBO could improve the precision of the transition probability estimates.

4.2.2 Long Sequence (826-bp)

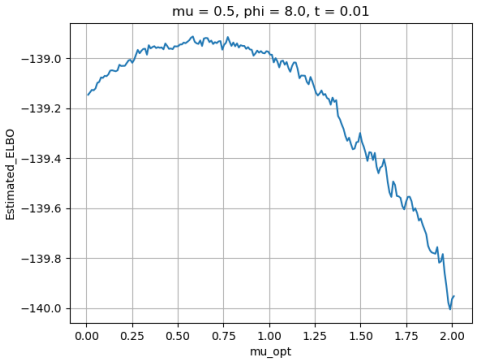
We now start to try a long sequence with 826-bp. We fix the mutation rate $\mu_{ori} = 0.5$, set the time $T = 0.01$ and use different levels of the context-dependence parameter ϕ_{ori} to generate evolutionary paths under different values of $\phi_{ori} = [2, 5, 8, 10]$. Figure 3 shows the relationship between estimated ELBO and μ under different values of the context-dependence parameter ϕ_{ori}. Note that we cannot compute the transition probability exactly because it requires an instantaneous rate matrix of dimension $4^{826} \times 4^{826}$, which is computationally infeasible.



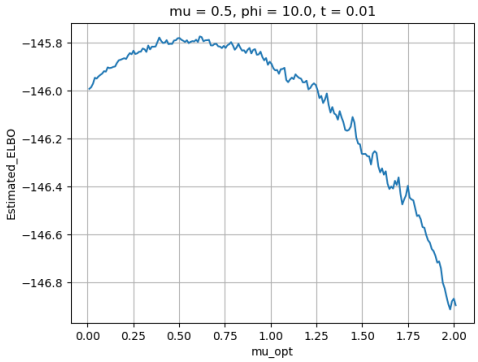
(a) $\phi_{ori} = 2$



(b) $\phi_{ori} = 5$



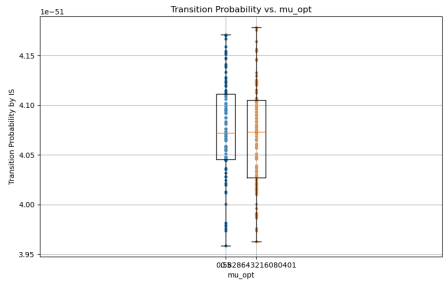
(c) $\phi_{ori} = 8$



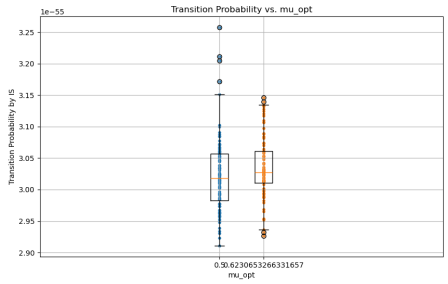
(d) $\phi_{ori} = 10$

Figure 3: Plots of ELBO vs. μ

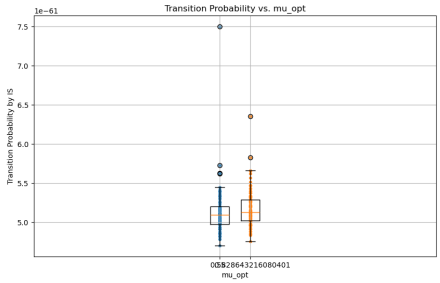
After determining the optimal parameter μ_{opt} that maximizes the estimated ELBO, we use it in the proposal distribution for importance sampling and compare its variability with that of the original parameter $\mu = 0.5$ in the CpG model. Specifically, with $N = 1000$, we sample mutation paths $\{P_i^{(j)}\}_{j=1}^{N}$ from the proposal distribution $q_{\mu_{\text{opt}}}(P_i^{(j)} \mid \mathbf{y}, \mathbf{x})$. These sampled paths are then used in the importance sampling procedure to estimate transition probabilities. We use boxplots to compare the variance of the estimates obtained with μ_{opt} against those from the original parameter $\mu_{ori} = 0.5$, which was not optimized via variational inference.



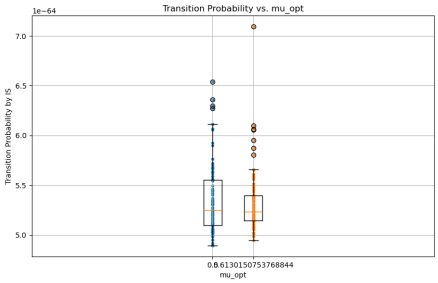
(a) $\phi_{ori} = 2$



(b) $\phi_{ori} = 5$



(c) $\phi_{ori} = 8$



(d) $\phi_{ori} = 10$

Figure 4: Boxplots of estimated transition probabilities for different ϕ_{ori} values

From the boxplots, we could see that using the optimized μ_{opt} increases the effective sample size (ESS) in importance sampling in context-dependent models to some extent.

However, maximizing the evidence lower bound (ELBO) for the CpG model is extremely time-consuming, as each ELBO evaluation relies on Monte Carlo integration, which requires a large number of samples to produce accurate estimates. Moreover, since the ELBO is estimated rather than computed exactly, this approximation can affect the stability of the optimization process. Instead, a direct grid search over μ values to be used as a candidates for μ_{opt} could be more computationally efficient. That said, for models with a large number of parameters—such as the ARMADiLLO model [11], which involves over 3,000 parameters—even grid search becomes computationally infeasible due to the high dimensionality of the parameter space.

5 Blockwise Importance Sampling

In this section, we introduce blockwise importance sampling as an alternative to optimize importance sampling.

5.1 CpG Model

Given the parameter for independent model: $\mu = 0.5$ is the standard rate for rate matrix, and ϕ is the adjusting parameter of μ with the new appearance of CG dinucleotides.

$$Q_{independent} = \begin{bmatrix} -3\mu & \mu & \mu & \mu \\ \mu & -3\mu & \mu & \mu \\ \mu & \mu & -3\mu & \mu \\ \mu & \mu & \mu & -3\mu \end{bmatrix}$$

Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$. Assume that the entire sequence can be evenly split into blocks of equal length. When we apply blockwise importance sampling into CpG model, the sequence is equally split into $\frac{n}{\mathrm{s}}$ number of non-overlapping partitions with s = block size. For each non-overlapping block, we use endpoint-conditioned CTMC sampling [7] to simulate the path independently. For context-dependent rate matrix Q_s for CpG model, we assign only one base-pair change index (like ATA $\rightarrow$ ATG) with the mutation rate μ, more than one base-pair change index (like TCA $\rightarrow$ CAG) with the mutation rate 0, and the one base-pair change index with the appearance of CG pair with the value (like ACC $\rightarrow$ ACG or CCA $\rightarrow$ CGA) with mutation rate $\mu\phi$, then impute the diagonal values by keeping the row sum to be 0. In this way, we could get the context-dependent rate matrix Q_s to do path sampling in a blockwise manner.

In this section, we consider $s \in \{1, 2, 3, 4, 5\}$, where s = 1 corresponds to the independent-site model. For $s = 1$, we set $N_1 = 1000$ and independently draw $P_1^{(1)}, \ldots, P_1^{(1000)}$ for each position using the rate matrix Q_1.

After recording the CPU time required to sample these 1000 paths, we compute how many paths $P_{\mathrm{s}}^{(i)}$ can be drawn within the same time budget for s = 2, 3, 4, and 5, respectively, and denote these quantities as $\{N_2, N_3, N_4, N_5\}$ (when s gets larger, the number of

paths that can be sampled gets smaller). Finally, we use N_s to draw paths $P_s^{(i)}$ using the rate matrix Q_s for $s \in \{2, 3, 4, 5\}$ respectively and use these paths in the blockwise importance sampling, where the likelihood is estimated under the blockwise proposal distribution $q_s(P^{(i)}|\mathbf{y}, \mathbf{x})$. Specifically, the unbiased estimator would be:

$$\widehat{\tilde{P}(\mathbf{y} \mid \mathbf{x})} = \frac{P_{Q_s}(\mathbf{y} \mid \mathbf{x})}{N_s} \sum_{i=1}^{N_s} \frac{\mathcal{L}(P_s^{(i)})}{q_s(P_s^{(i)})}$$

For each block $b = 1, 2, ..., B$ with $B = \frac{n}{s}$, let $x^{(b)}, y^{(b)} \in \{A, G, C, T\}^s$ be the block of the whole sequences $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Then,

$$P_{Q_s}(\mathbf{y} \mid \mathbf{x}) = \prod_{b=1}^{B} (e^{TQ_s})_{x^{(b)}, y^{(b)}}.$$

With given $\mu = 0.5$ and $\phi = 0.3$, we use the forward simulation method [7] to simulate the end sequence $\mathbf{y}_T$ from start sequence $\mathbf{x}$ (360 bp) with time T = 0.05, 0.1 and 0.2 respectively and get $\mathbf{y}_{0.05}, \mathbf{y}_{0.1}, \mathbf{y}_{0.2}$ respectively (the mutation numbers of $\mathbf{x} \to \mathbf{y}_{0.05}$, $\mathbf{x} \to \mathbf{y}_{0.1}$, and $\mathbf{x} \to \mathbf{y}_{0.2}$ are 23, 56, and 98, respectively).

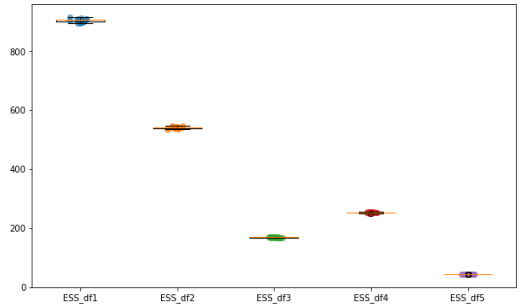
For each sequence $\mathbf{y}$, we change time into $T_{0.05}, T_{0.1}, T_{0.2}, T_{0.5}, T_{1.0}$ respectively and test if blockwise sampling could both enhance the ESS and be efficient. In Figures 5-7, we plot the ESS for each case, with s increasing sequentially from 1 (left) to 3 (right).

With N_s weighted samples with normalized importance weights w_1, w_2, ..., w_{N_s}, we have:

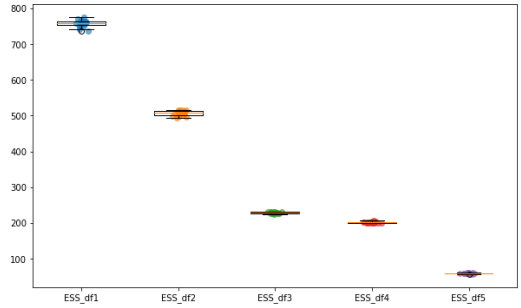
$$\mathcal{W}_s(P^{(i)}) = \frac{\mathcal{L}(P_s^{(i)})}{q_s(P_s^{(i)})}$$

$$w_i = \frac{\mathcal{W}_s(P^{(i)})}{\sum_{j=1}^{N} \mathcal{W}_s(P^{(j)})}$$

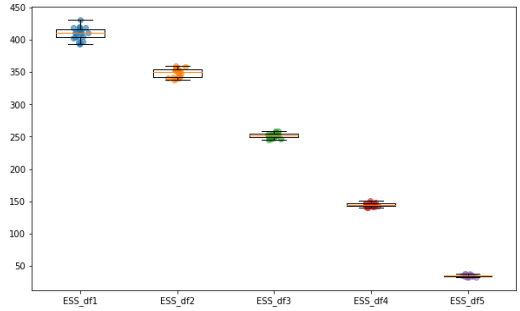
$$\text{ESS}_s = \frac{1}{\sum_{i=1}^{N} w_i^2}.$$



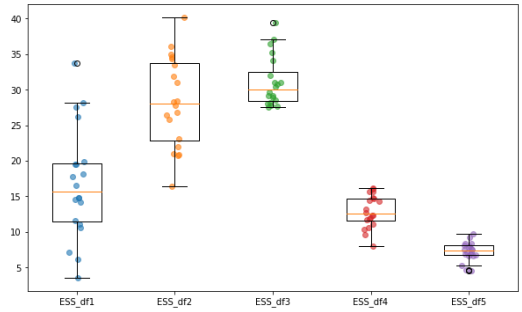
(a) T = 0.05



(b) T = 0.1

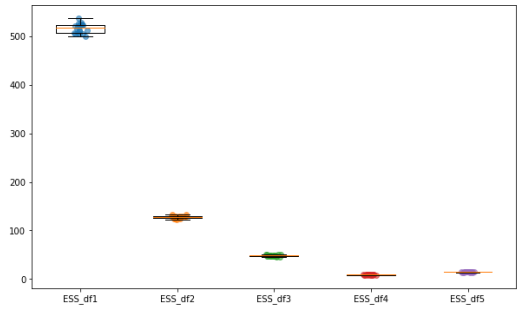


(c) T = 0.2

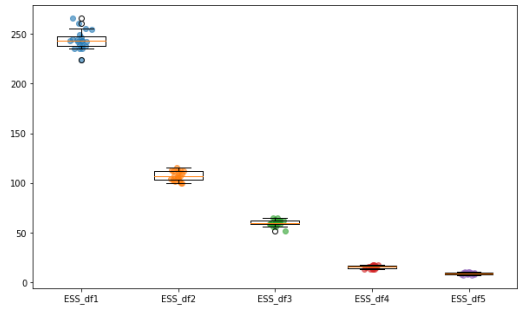


(d) T = 0.5

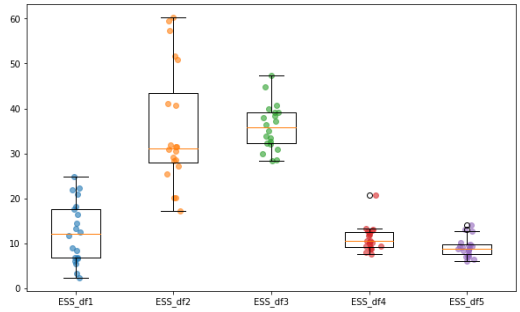
Figure 5: Plots of Effective Sample Size (ESS) for $\mathbf{x} \to \mathbf{y}_{0.05}$ (23 mutations)



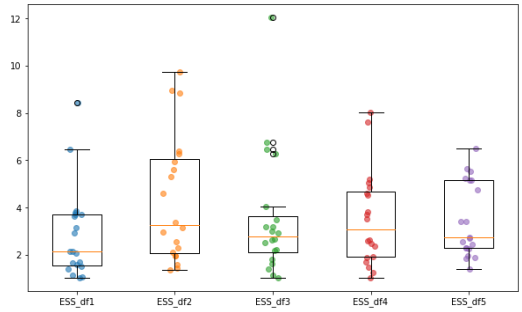
(a) T = 0.1



(b) T = 0.2

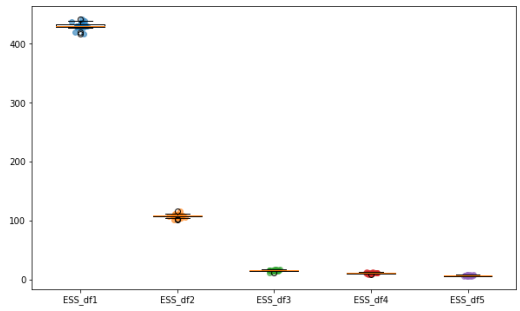


(c) T = 0.5

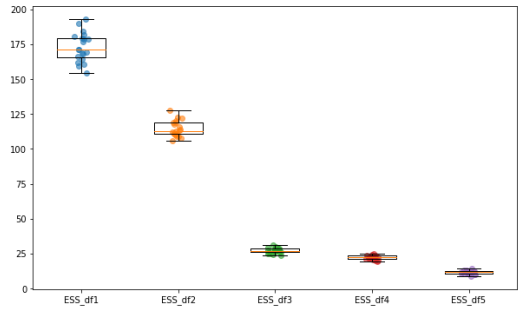


(d) T = 1.0

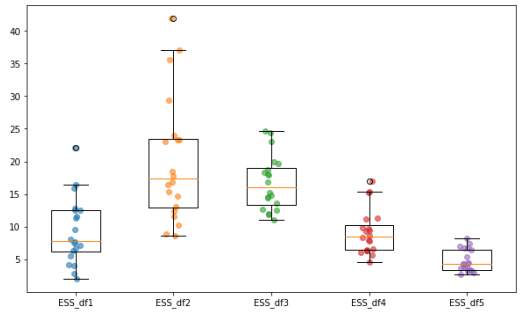
Figure 6: Plots of Effective Sample Size (ESS) for $\mathbf{x} \to \mathbf{y}_{0.1}$ (56 mutations)



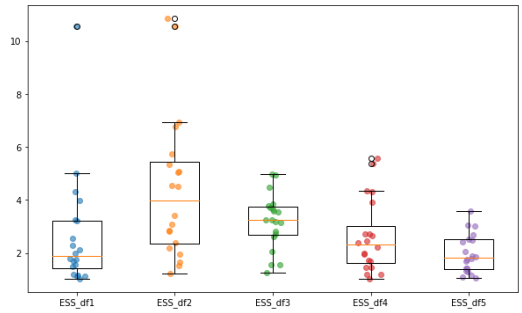
(a) T = 0.1



(b) T = 0.2



(c) T = 0.5



(d) T = 1.0

Figure 7: Plots of Effective Sample Size (ESS) for $\mathbf{x} \to \mathbf{y}_{0.2}$ (98 mutations)

From figures 5-7, we observe that when the endpoint sequences $\mathbf{x}$ and $\mathbf{y}$ are fixed, increasing the total mutation time T leads to improved performance of blockwise sampling. In particular, blockwise sampling with s = 3 starts exhibiting a higher ESS and lower variability in estimating transition probability compare to s = 1 and 2, especially when T = 0.5. However, when T is too large (T = 1.0), the ESS would be extremely low for all cases so that the blockwise won't be efficient in further case.

This occurs because when the time interval is small (e.g., T = 0.1), blockwise modeling provides limited improvement to importance sampling, as it allows only 1–2 mutations to occur at each position. However, as the time interval increases (e.g., T = 0.5), more mutations are allowed to occur, making the blockwise approach better capture dependencies between neighboring sites and improve the precision of estimates.

5.2 ARMADiLLO Model

For ARMADiLLO model [11], the experimentation setup for blockwise sampling is similar, except the fact that we should use a different design of rate matrices for each position

of sequence. Specifically, for each position x_i, the rate matrix Q depends on the 4-mer neighborhood surrounding that center position (i.e.,$(x_{i-2}, x_{i-1}, x_{i+1}, x_{i+2})$). As a result, we have $4^4 = 256$ unique rate matrices for ARMADiLLO model. For the ARMADiLLO case, the ESS is computed in the same way as in the CpG case in Section 5.1.

5.2.1 Full Blockwise IS for ARMADiLLO Model

For $s = 1$, we assign a unique rate matrix $Q_1^{(b)}$ to each position $b = 1, \ldots, B$ of the sequence $\mathbf{x}$ or $\mathbf{y}$ with $B = \frac{n}{s}$, which is used to simulate paths under the independent-site model.

In the blockwise setting with s = 2, 3, 4 and 5, we construct the matrices $Q_2^{(b)}$, $Q_3^{(b)}$, $Q_4^{(b)}$, and $Q_5^{(b)}$ by directly projecting mutation rate values from the Q_1 matrix into $4^{\mathrm{s}} \times 4^{\mathrm{s}}$ matrix. In particular, indices of Q_s with single-nucleotide substitution are filled using the mutation rates from Q_1, while indices with multiple simultaneous mutations are set to zero. The diagonal indices are finally imputed to ensure that each row sums to zero. Note that each Q_s also depends on the 4-mer neighborhood surrounding the entire block.

With the same method in section 5.1, we get N_s for each $s \in \{2, 3, 4, 5\}$. For each s, we could do path simulations of $P_s^{(1)}$, $P_s^{(2)}$, ..., $P_s^{(N_s)}$. Then we use these paths to estimate the transition probabilities of context-dependent models:

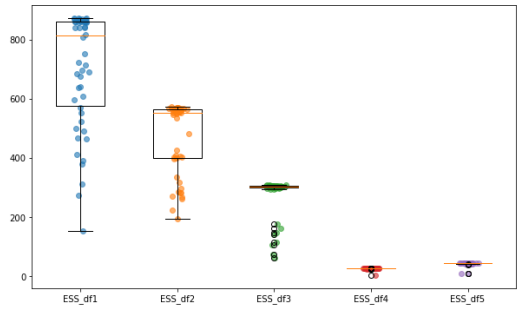
$$\widehat{\tilde{P}(\mathbf{y} \mid \mathbf{x})} = \frac{P_{Q_s}(\mathbf{y} \mid \mathbf{x})}{N_s} \sum_{i=1}^{N_s} \frac{\mathcal{L}(P_s^{(i)})}{q_s(P_s^{(i)})}$$

For each block $b = 1, 2, ..., B$ with $B = \frac{n}{s}$, let $x^{(b)}, y^{(b)} \in \{A, G, C, T\}^s$ be the block of the whole sequences $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Then,

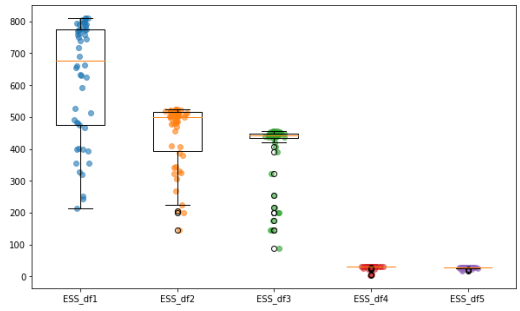
$$P_{Q_s}(\mathbf{y} \mid \mathbf{x}) = \prod_{b=1}^{B} (e^{TQ_s^{(b)}})_{x^{(b)}, y^{(b)}}.$$

With the same setup in CpG model, we use the forward simulation method [7] to simulate the end sequence $\mathbf{y}_T$ from start sequence $\mathbf{x}$ (360 bp) with time T = 0.05, 0.1, and 0.2 respectively and get $\mathbf{y}_{0.05}$, $\mathbf{y}_{0.1}$, and $\mathbf{y}_{0.2}$ respectively. (the mutation numbers of $\mathbf{x} \to \mathbf{y}_{0.05}$, $\mathbf{x} \to \mathbf{y}_{0.1}$ and $\mathbf{x} \to \mathbf{y}_{0.2}$ are 17, 34 and 62, respectively)

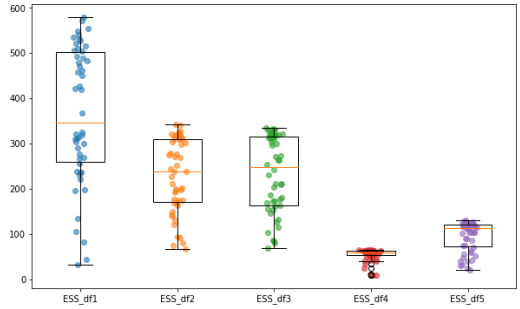
For each type of $\mathbf{y}_T$, we change time into $T_{0.05}, T_{0.1}, T_{0.2}, T_{0.5}, T_{1.0}$ respectively and test if blockwise sampling could both enhance the ESS and be efficient. In Figures 8–10, we plot the ESS for each case, with s increasing sequentially from 1 (left) to 5 (right).



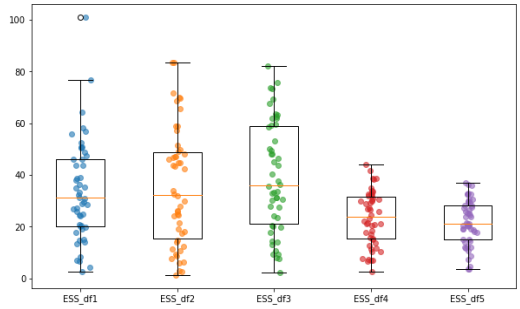
(a) T = 0.05



(b) T = 0.1

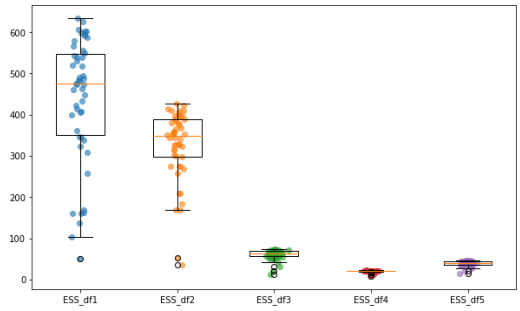


(c) T = 0.2

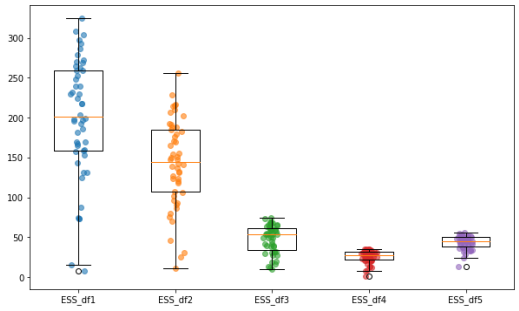


(d) T = 0.5

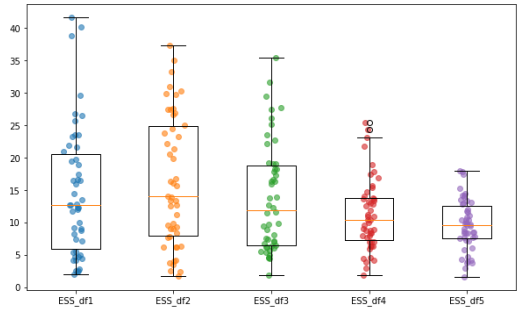
Figure 8: Plots of Effective Sample Size (ESS) for $\mathbf{x} \to \mathbf{y}_{0.05}$ (17 mutations)



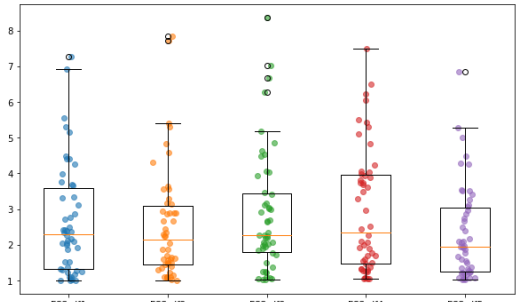
(a) T = 0.1



(b) T = 0.2

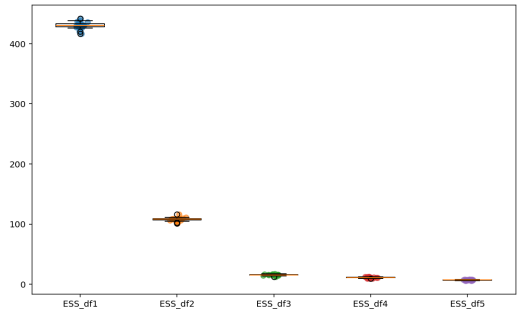


(c) T = 0.5

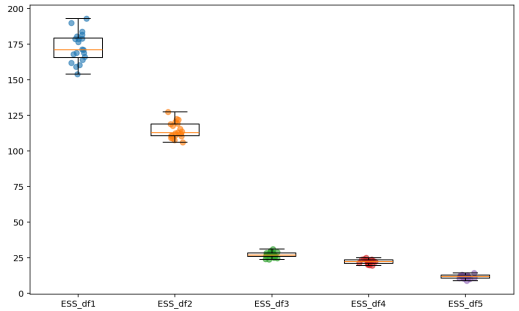


(d) T = 1.0

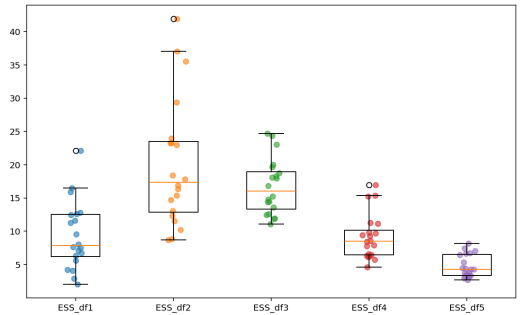
Figure 9: Plots of Effective Sample Size (ESS) for $\mathbf{x} \to \mathbf{y}_{0.1}$ (34 mutations)



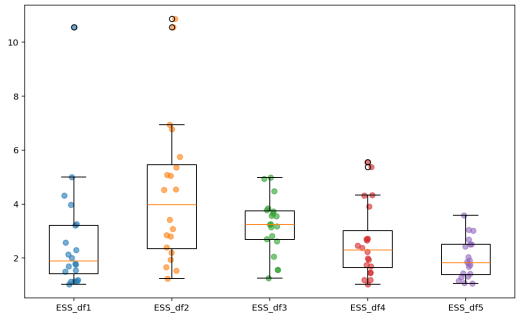
(a) T = 0.1



(b) T = 0.2



(c) T = 0.5



(d) T = 1.0

Figure 10: Plots of Effective Sample Size (ESS) for $\mathbf{x} \to \mathbf{y}_{0.2}$ (62 mutations)

From the plots, we observe that even as the number of mutations from sequence $\mathbf{x}$ to $\mathbf{y}$ increases, blockwise sampling still does not outperform the independent-site model in terms of efficiency. It demonstrates that full blockwise sampling for is not efficient enough for ARMADiLLO model.

5.2.2 Selective Blockwise IS for ARMADiLLO Model

In the previous section, we evenly split the whole sequence into blocks, which does not appear to be efficient. We then attempt to selectively use blockwise path sampling at specific positions using $Q_s^{(b)}$, while sampling rest of the sequence independently using $Q_1^{(j)}$ $(j \neq b)$.

Specifically, when the exit rate at a particular position i for sequence $\mathbf{x}$ (i.e., $-Q_1[x_b, x_b]$) exceeds a predefined threshold, the position i is identified as a high exit rate site, and a block of fixed size (s = 5 for the ARMADiLLO model) is formed centered at that site for path sampling using $Q_5^{(i)}$. For other sites $j \neq b$, use $Q_1^{(j)}$ to do path sampling.

For the ARMADiLLO model, we consider a 360-bp sequence with a time T = 0.2 and 50 endpoint mutations from $\mathbf{x}$ to $\mathbf{y}$. We set the exit rate threshold to 3.5; for any base with an exit rate exceeding this threshold, we form a block (with s = 5) centered around it. For each estimate by IS, we compute its corresponding $\text{ESS}_{selective}$ and compare it with that of independent sampling, ESS_1. In figure 11, we can see that selective blockwise importance sampling (right) outperforms independent importance sampling (left), as it achieves a significantly higher ESS.

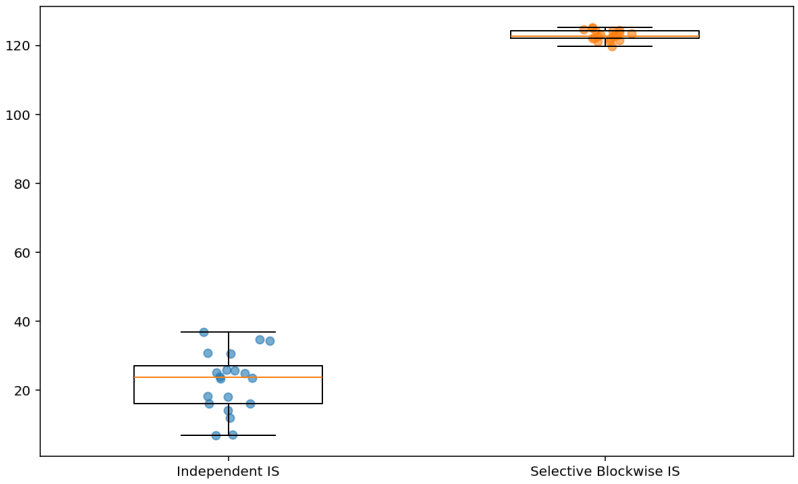


Figure 11: Plots of Effective Sample Size (ESS) for 360-bp sequence with selective blockwise importance sampling

However, the empirical value of threshold requires theoretical justification and more experimental to be justified. In addition, the specific example involves a sparse bases with a high exit rate within the sequence. When the bases with high exit rates are more densely distributed across the start sequence, it becomes more challenging to select these non-overlapping blocks in importance sampling. Block selection in such extreme cases would be an interesting direction for future research.

6 Parameter Estimation

In this section we are going through parameter estimation. Specifically, we are going to use IS-EM to estimate the free parameters for both CpG model [9] and ARMADiLLO model [11].

Assume we are given M observed sequence pairs $(\mathbf{x}_i, \mathbf{y}_i)$ for $i = 1, ..., M$ with an unknown parameter vector $\boldsymbol{\theta}$. Each observed sequence $\mathbf{y}_i$ is assumed to be arise from some unknown latent path $\{P_i^{(j)}\}_{j=1}^N$. We are going to use IS-EM (Importance Sampling-Expectation-Maximization) algorithm to estimate $\boldsymbol{\theta}$.

6.1 E-Step: Importance Weights

Initialize the unknown parameter $\boldsymbol{\theta}^{(0)}$. For each iteration t:

1. Sample N paths $\{P_i^{(j)}\}_{j=1}^N$ from a proposal distribution $q(P_i^{(j)}|\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\theta}^{(t)})$.

2. Compute importance weights for each sample:

$$w_i^{(j)} = \frac{\mathcal{L}(P_i^{(j)}|\boldsymbol{\theta}^{(t)})}{q(P_i^{(j)}|\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\theta}^{(t)})}$$

3. Estimate the expected complete log-likelihood with M pairs of $(\mathbf{x}_i, \mathbf{y}_i)$:

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) \approx \sum_{i=1}^{M} \sum_{j=1}^{N} \tilde{w}_i^{(j)} \log \mathcal{L}(P_i^{(j)}|\boldsymbol{\theta})$$

$$\tilde{w}_i^{(j)} = \frac{w_i^{(j)}}{\sum_{j=1}^N w_i^{(j)}}$$

6.2 M-Step: Gradient-Based Parameter Update

1. Compute the weighted gradient:

$$\nabla_\theta Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = \sum_{i=1}^{M} \sum_{j=1}^{N} \tilde{w}_i^{(j)} \nabla_\theta \log \mathcal{L}(P_i^{(j)}|\boldsymbol{\theta}^{(t)})$$

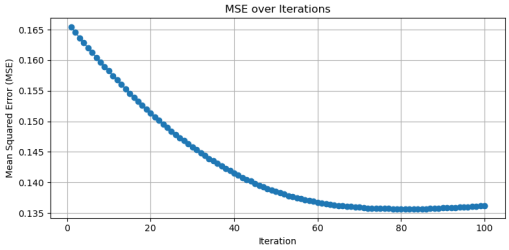
2. For all $\theta^{(t)} \in \boldsymbol{\theta}^{(t)}$, update $\theta^{(t+1)}$ from $\theta^{(t)}$ using gradient ascent:

$$\theta^{(t+1)} = \theta^{(t)} + \eta \sum_{i=1}^{M} \sum_{j=1}^{N} \tilde{w}_i^{(j)} \nabla_\theta \log \mathcal{L}(P_i^{(j)} | \boldsymbol{\theta}^{(t)})$$

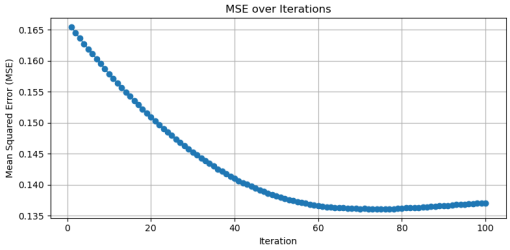
We initialize the step size parameter $\eta = 0.1$. During the whole process, if any computed element $\theta \in \boldsymbol{\theta}$ becomes negative, we reduce the step size by multiplying η by 0.1. This adaptive adjustment helps maintain numerical stability and ensures that all mutation rates $\theta \in \boldsymbol{\theta}$ remain valid (i.e., non-negative). The algorithm iterates until the parameters θ converge.

6.3 Parameter Estimation for ARMADiLLO Model

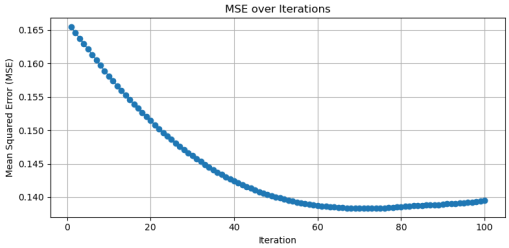
First, we perform parameter estimation for the ARMADiLLO model [11]. For $i = 1, \ldots, 10$, since the ARMADiLLO model has no closed-form stationary distribution, we generate each sequence $\mathbf{x}_i$ of length K by sampling each position x_k independently from a uniform distribution over $[A, C, T, G]$, with probabilities $[0.25, 0.25, 0.25, 0.25]$ for $k = 1, \ldots, K$. Then, each corresponding $\mathbf{y}_i$ is simulated from the ARMADiLLO model given $\mathbf{x}_i$ and time $T = 0.2$, using the true parameter $\boldsymbol{\theta}$. Given all $(\mathbf{x}_i, \mathbf{y}_i)$ pair, assume we don't know the real parameters $\boldsymbol{\theta}$, we use IS-EM to estimate $\boldsymbol{\theta}$ in an iterative way.



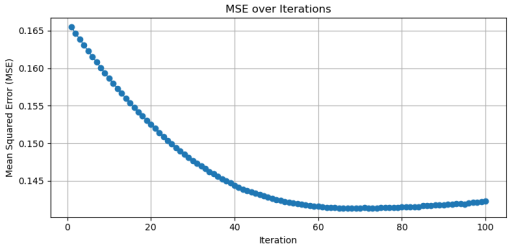
(a) 100-bp, T = 0.2, M = 10, N = 100



(b) 100-bp, T = 0.2, M = 10, N = 100



(c) 100-bp, T = 0.2, M = 10, N = 1000



(d) 100-bp, T = 0.2, M = 10, N = 1000

Figure 12: Plots of MSE with 100-bp, T = 0.2 with 10 pairs of ($\mathbf{x}_i$, $\mathbf{y}_i$) in ARMADiLLO Model

For the ARMADiLLO model, we can see that the MSE for all parameters decreases over the iterations to some extent, but the improvement is limited. This is because the simulated mutation path for the observed $(\mathbf{x}_i, \mathbf{y}_i)$ pair can't fully capture all of the more than 3,000 parameters, so many parameters don't get optimized over the iterations.

Besides the accuracy, another limitation of this method is its computational cost. In our experiments for ARMADiLLO model, running the algorithm on just 10 simulated sequences of length 100 base pairs took several days to complete. In reality, DNA sequences are more than thousands of base pairs long, making it computationally infeasible to estimate over 3000 parameters through iterative updates becomes highly challenging under these constraints.

6.4 Parameter Estimation in K80 + CpG Model

We also do the parameter estimation in K80 + CpG model [9, 13], which differs from the CpG model described in Chapters 4 and 5. For the triplet of adjacent nucleotides (x_{i-1}, x_i, x_{i+1}), the substitution rate from the base x_i to base b is defined as:

$$\gamma(b; x_{i-1}, x_i, x_{i+1}) = \begin{cases} \frac{Q[x_i, b]}{\lambda}, & \text{if } (x_{i-1}, x_i) = (C, G) \text{ or } (x_i, x_{i+1}) = (C, G). \\ Q[x_i, b], & \text{otherwise.} \end{cases}$$

With the rate of transition α and the rate of transversion β, the rate matrix for the independent K80 model is given by:

$$Q = \begin{bmatrix} -\alpha - 2\beta & \alpha & \beta & \beta \\ \alpha & -\alpha - 2\beta & \beta & \beta \\ \beta & \beta & -\alpha - 2\beta & \alpha \\ \beta & \beta & \alpha & -\alpha - 2\beta \end{bmatrix}$$

Assuming $\boldsymbol{\theta} = (\alpha, \beta, \lambda)$ and N = 4. With i = 1, ..., 4, we simulate $\mathbf{x}_i$ from stationary distribution of CpG model [9], then simulate each $\mathbf{y}_i$ from K80 + CpG model given its corresponding $\mathbf{x}_i$ with time $T \in [0.1, 0.2, 0.3]$ and true parameter $(\alpha, \beta, \lambda) = (0.4, 0.2, 0.15)$. Assume we are given observed $(\mathbf{x}_i, \mathbf{y}_i)$ sequence pairs without knowing the true parameters (α, β, λ), we use IS-EM to estimate $(\hat{\alpha}, \hat{\beta}, \hat{\lambda})$ iteratively. For each $\mathbf{x}_i$, with the stationary distribution $P_\lambda(\mathbf{x}_i)$ [9]:

1. **Estimate $\hat{\lambda}$:**

$$\hat{\lambda} = \arg\max_{\lambda} \sum_{i=1}^{N} \log P_\lambda(\mathbf{x}_i)$$

2. **Estimate $\hat{\alpha}$ and $\hat{\beta}$ via IS-EM:**

   Given the estimated $\hat{\lambda}$ from step 1, initialize $\alpha^{(0)}$ and $\beta^{(0)}$. For $t = 1, 2, ..., 100$:

$$\alpha^{(t)} = \alpha^{(t-1)} + \eta \sum_{i=1}^{M} \sum_{j=1}^{N} \tilde{w}_i^{(j)} \nabla_\alpha \log \mathcal{L}\big(P_i^{(j)} \mid \alpha^{(t)}, \beta^{(t)}, \hat{\lambda}\big)$$

$$\beta^{(t)} = \beta^{(t-1)} + \eta \sum_{i=1}^{M} \sum_{j=1}^{N} \tilde{w}_i^{(j)} \nabla_\beta \log \mathcal{L}\big(P_i^{(j)} \mid \alpha^{(t)}, \beta^{(t)}, \hat{\lambda}\big)$$

   Note that the complete-data log-likelihood expression $\log \mathcal{L}(P_i^{(j)} \mid \alpha^{(t)}, \beta^{(t)}, \hat{\lambda})$ is the same as the one used in the MCMC-EM algorithm [9] and MCMC-GEM [13].

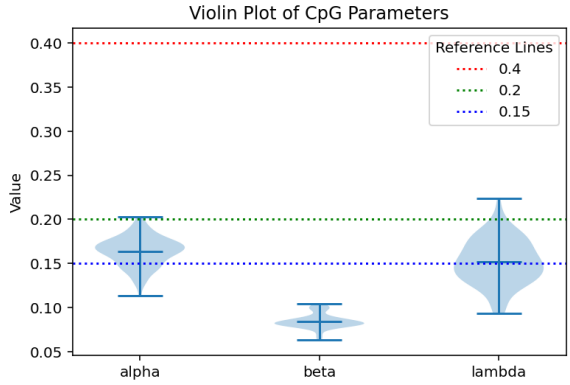
3. **Convergence:**

   Repeat the iterations until $(\alpha^{(t)}, \beta^{(t)})$ converge. Define our final estimates as $(\hat{\alpha}, \hat{\beta})$.

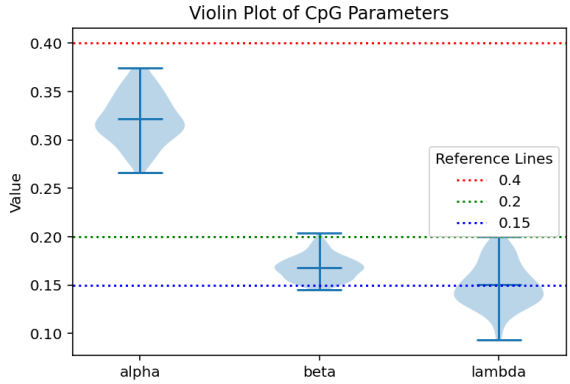
In the simulation study, we generate 4 pairs of $(\mathbf{x}_i, \mathbf{y}_i)$ for each trial with different values of T ($T \in \{0.1, 0.2, 0.3\}$), perform 100 trials in total, and get 100 estimates of $(\hat{\alpha}, \hat{\beta}, \hat{\lambda})$ as described in [13]. These estimates are visualized in violin plots (Figure 13) and compared against the true parameter values $(\alpha, \beta, \lambda) = (0.4, 0.2, 0.15)$. Table 1 reports the average estimates based on 100 trials.

T	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\lambda}$
0.1	0.163	0.084	0.152
0.2	0.322	0.168	0.150
0.3	0.553	0.269	0.152

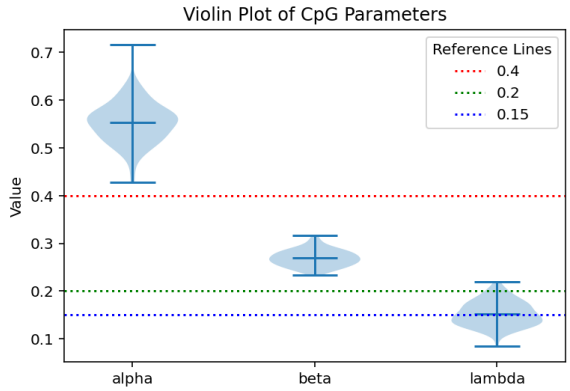
Table 1: Mean of Estimates with Different T values



(a) T = 0.1



(b) T = 0.2



(c) T = 0.3

Figure 13: Violin plot of CpG parameters (α, β, λ) with dotted reference lines at 0.4, 0.2, and 0.15.

From the plots we could see that only the parameter λ approximates the true value. Because we have the expression of stationary distribution of $\mathbf{x}_i$ for i = 1, ..., N, we don't need to estimate λ through importance sampling. However, compare to MCMC-GEM [13] , the IS-EM for K80 + CpG model is not accurate enough for estimating (α, β) pair. It reflects that while IS-EM is computational efficient, it does not seem to outperform MCMC-GEM in terms of accuracy. One of the potential reason is that our proposal distribution, based on model K80, doesn't capture the dependencies introduced by λ. As a result, it would lead to a high variance in importance weights and unstable estimation of parameters (α, β). It is possible that improving the proposal distribution would help reduce variance and improve estimation accuracy.

7 Conclusion

Importance sampling offers a practical approach for inference in context-dependent models, particularly when likelihood evaluations are intractable or expensive. However, it is always limited by high variance and the need for a large number of samples, especially when the proposal distribution is not close to the target distribution. As demonstrated in this thesis, although blockwise importance sampling and variational inference could improve performance to some extent, these improvements often come at the cost of increased computational time. In addition, choosing a proposal distribution with low variance would be beneficial for improving the accuracy of parameter estimation.

In conclusion, optimizing importance sampling for context-dependent evolutionary models remains an open and active area of research. More computationally efficient ways are expected in future studies.

Bibliography

[1] Mathews, Joseph (2025). Advances in Sequential Monte Carlo Methods and Site-Dependent DNA Evolution Models. PhD thesis, Department of Statistical Science, Duke University.

[2] Mathews, J. and Schmidler, S.C. (2025). Approximating Marginal Likelihoods in Evolutionary Models Under Site-Dependence. (submitted)

[3] Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. Journal of the American Statistical Association 112.518 (2017): 859-877.

[4] Murphy, Kevin P. Machine learning: a probabilistic perspective. MIT Press, 2012.

[5] Bryant, David, Nicolas Galtier, and Marie-Anne Poursat. Likelihood calculation in molecular phylogenetics. (2007): 33-62.

[6] Chen, Ming-Hui, Lynn Kuo, and Paul O. Lewis, eds. Bayesian phylogenetics: methods, algorithms, and applications. CRC Press, 2014.

[7] Hobolth, Asger, and Eric A. Stone. Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. The Annals of Applied Statistics 3.3 (2009): 1204.

[8] Mathews, Joseph, et al. Computing the inducibility of B cell lineages under a context-dependent model of affinity maturation: Applications to sequential vaccine design. bioRxiv (2023).

[9] Hobolth, Asger. A Markov chain Monte Carlo expectation maximization algorithm for statistical analysis of DNA sequence evolution with neighbor-dependent substitution rates. Journal of Computational and Graphical Statistics 17.1 (2008): 138-162.

[10] Jensen, J., and A.-M. Pedersen. Probabilistic models of DNA sequence evolution with context-dependent rates of substitution. Advances in Applied Probability 32.2 (2000): 499–517.

[11] Wiehe, Kevin, et al. Functional relevance of improbable antibody mutations for HIV broadly neutralizing antibody development. Cell Host & Microbe 23.6 (2018): 759-765.

[12] Jukes, T. H., Cantor, C. R., et al. Evolution of protein molecules. In: H. N. Munro (Ed.), Mammalian Protein Metabolism, vol. 3, pp. 21–132, Academic Press, 1969.

[13] Li, Y., Mathews, J., and Schmidler, S.C. (2025). On Gibbs sampling for endpoint-conditioned neighbor-dependent sequence evolution models. (submitted)