Subsampling Sequential Monte Carlo for Phylogenetic 'Tall Data'

Shijia Wang[1], Liangliang Wang[1], Alexandre Bouchard-Côté[2]

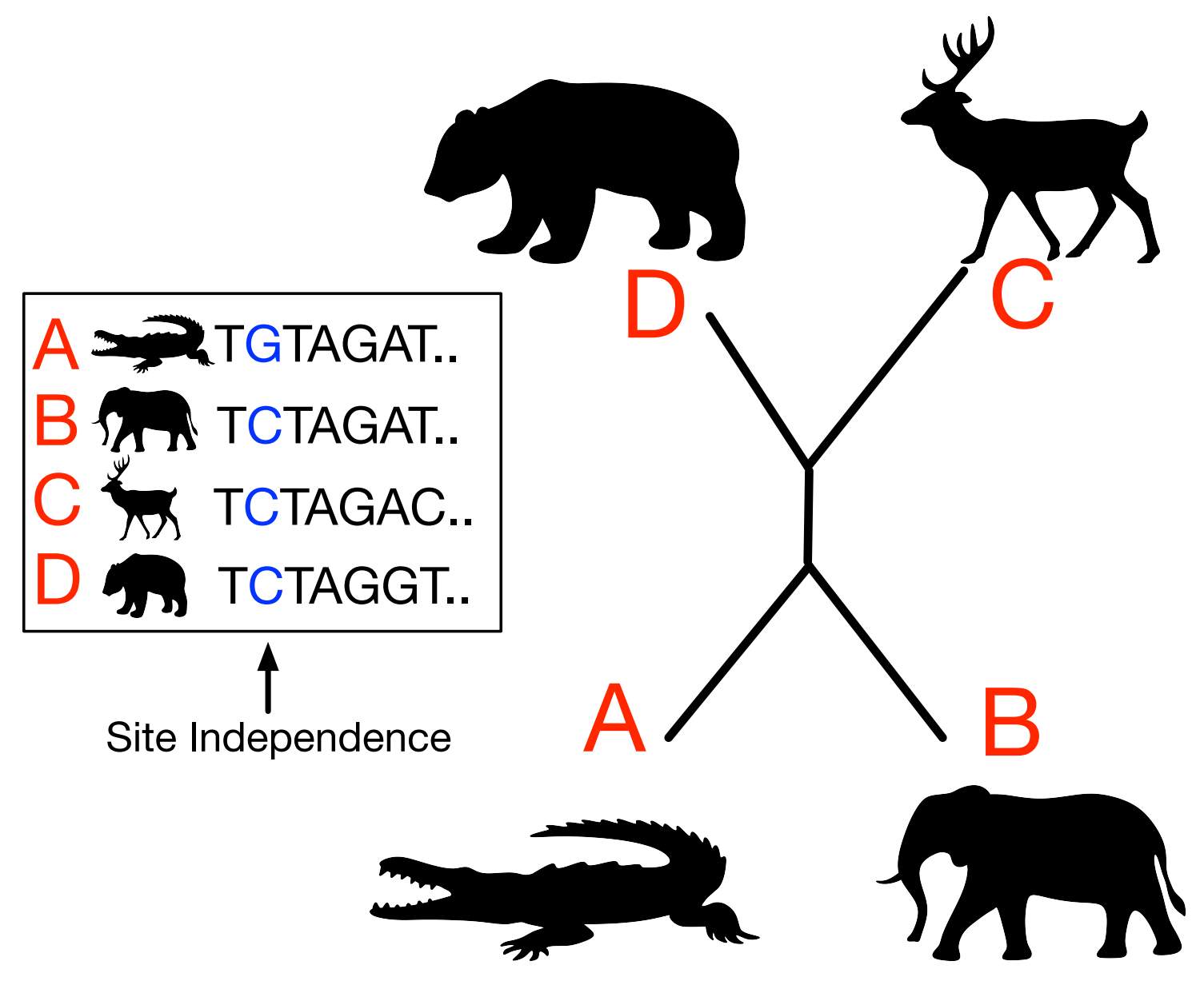
[1]Department of Statistics and Actuarial Science, Simon Fraser University, Canada
[2]Department of Statistics, University of British Columbia, Canada

Abstract

We propose a novel method for Bayesian phylogenetic 'tall data' inference by combing the idea of subsampling with annealed sequential Monte Carlo (SMC) [1]. Unlike the previous SMC methods in phylogenetics, the subsampling SMC has the same state space for all the intermediate distributions, which allows standard Markov chain Monte Carlo (MCMC) tree moves to be utilized as the basis for SMC proposal distributions. The proposed algorithm possesses the attractive property of SMC methods, as well as the ability of subsampling [2] for 'tall data' inference.

Bayesian 'Big Data' Phylogenetics

- We are interested in the study of evolutionary relationships among biological species (taxa).



- Data y: Super long biological sequences (e.g. DNA sequence) of a set of species.
- The likelihood model is based on site independence assumption.
- Our interest is the posterior distribution of parameters $x = (\theta, t)$

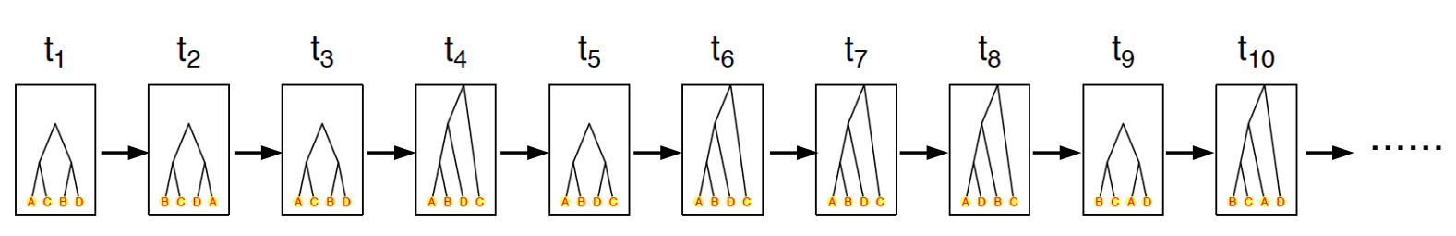
$$\pi(x) = \frac{P(y|x)p(x)}{P(y)}.$$

- Posterior expectation of $E(\varphi(x)) = \int \pi(dx)\varphi(x)$.

Challenges in inference

- Expensive in evaluating the unnormalized posterior.
- Multimodal posterior distributions.
- The estimation of normalization constant $P(y) = \int P(y|x)p(x)dx$.

Markov chain Monte Carlo



- Cannot fully explore the multimodal tree space.
- Not scalable to 'big data'.
- Challenge in normalization constant estimation.

Intermediate Sequence of Subsampling SMC

1. Let us decompose the unnormalized posterior distribution as

$$\gamma(x) = p(x) \prod_{s=1}^{\#S} p(y_s|x),$$

   - s refers to one batch of sites in biological sequence.
   - $\#S$ represents the total number of batches.

2. Specify a power sequence $0 = \phi_0 < \phi_1 < \cdots < \phi_R = 1$.
3. Define the sequence of intermediate distributions for subsampling as follows:

$$\gamma_{\phi_r}(x) = p(x) \prod_{s=1}^{\#S} p(y_s|x)^{\psi(s,\phi_r)},$$

where

$$\psi(s, \phi_r) = \begin{cases} 1 & \text{if } \phi_r \geq s/\#S, \\ 0 & \text{if } \phi_r \leq (s-1)/\#S, \\ \#S \cdot \phi_r - (s-1) & \text{otherwise.} \end{cases}$$

4. It's a general version of the target sequence for annealed SMC algorithm[1].

Subsampling SMC Algorithmic Framework

- Figure 1 provides an overview of the subsampling SMC framework.
- Use a list of weighted samples $(x_{r,k}, W_{r,k})_{k=1}^{K}$ to approximate $\pi_{\phi_r}(x)$.
- The algorithm alternates between the following three steps:
  1. compute the weights using samples from the previous iteration,
  2. perform MCMC moves to propose samples,
  3. resample to prune particles with small weights (triggered by effective sample size).
- We perform weighting before proposing new samples, which is different from the standard SMC algorithm [3].
- The fact that $W_{r,k}$ only depends on $x_{r-1,k}$ allows us to adaptively determine $\psi(s, \phi_r)$ $(r = 1, \ldots, R)$.
- Output an approximation of posterior $\pi(x) \approx \sum_k W_{R,k}\delta_{\tilde{x}_{R,k}}(x)$ and an approximation of normalizing constant

$$\hat{Z}_K := \prod_{r=1}^{R} \frac{1}{K} \sum_{k=1}^{K} w_{r,k}.$$

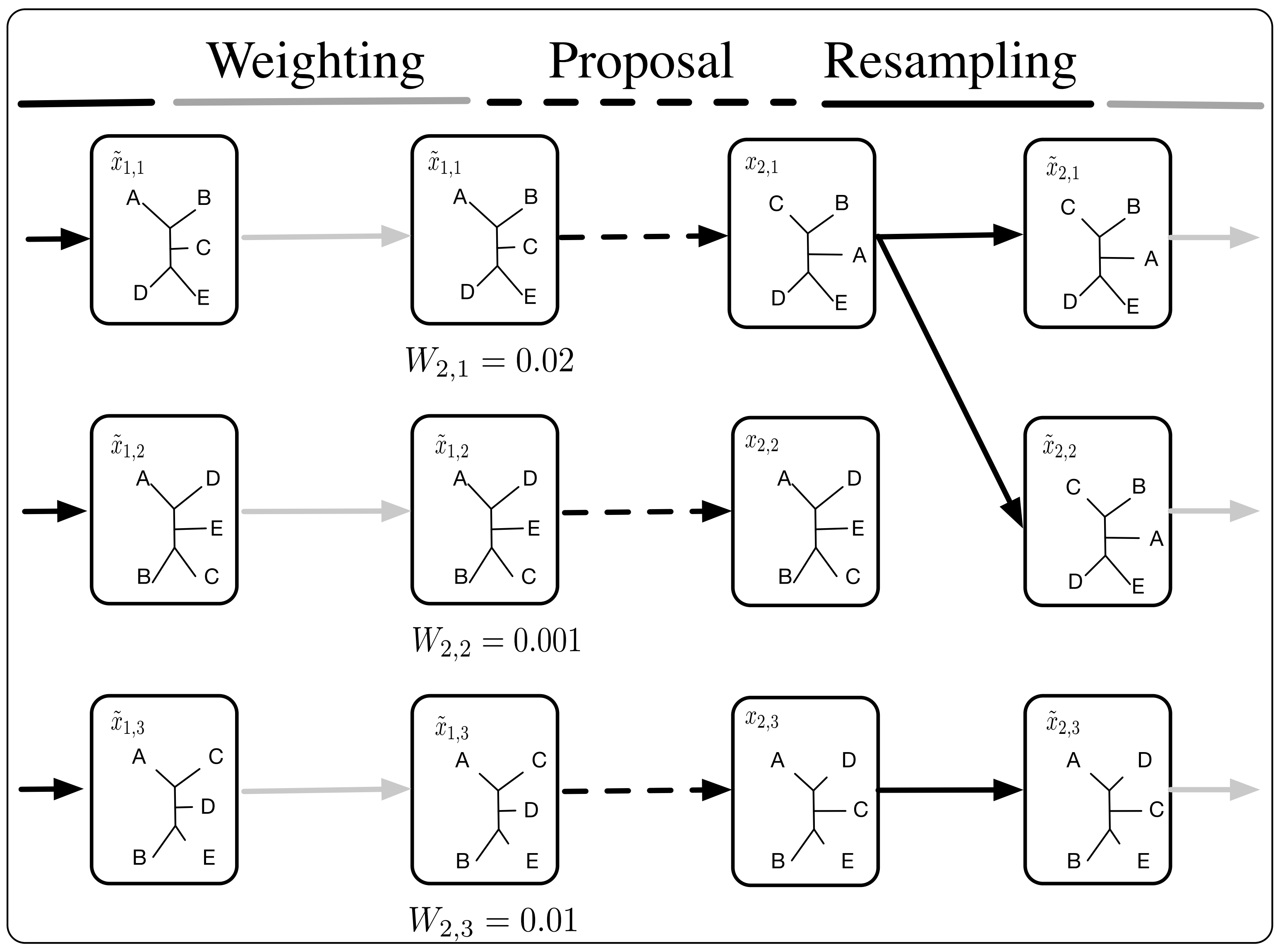


Figure 1: Subsampling SMC

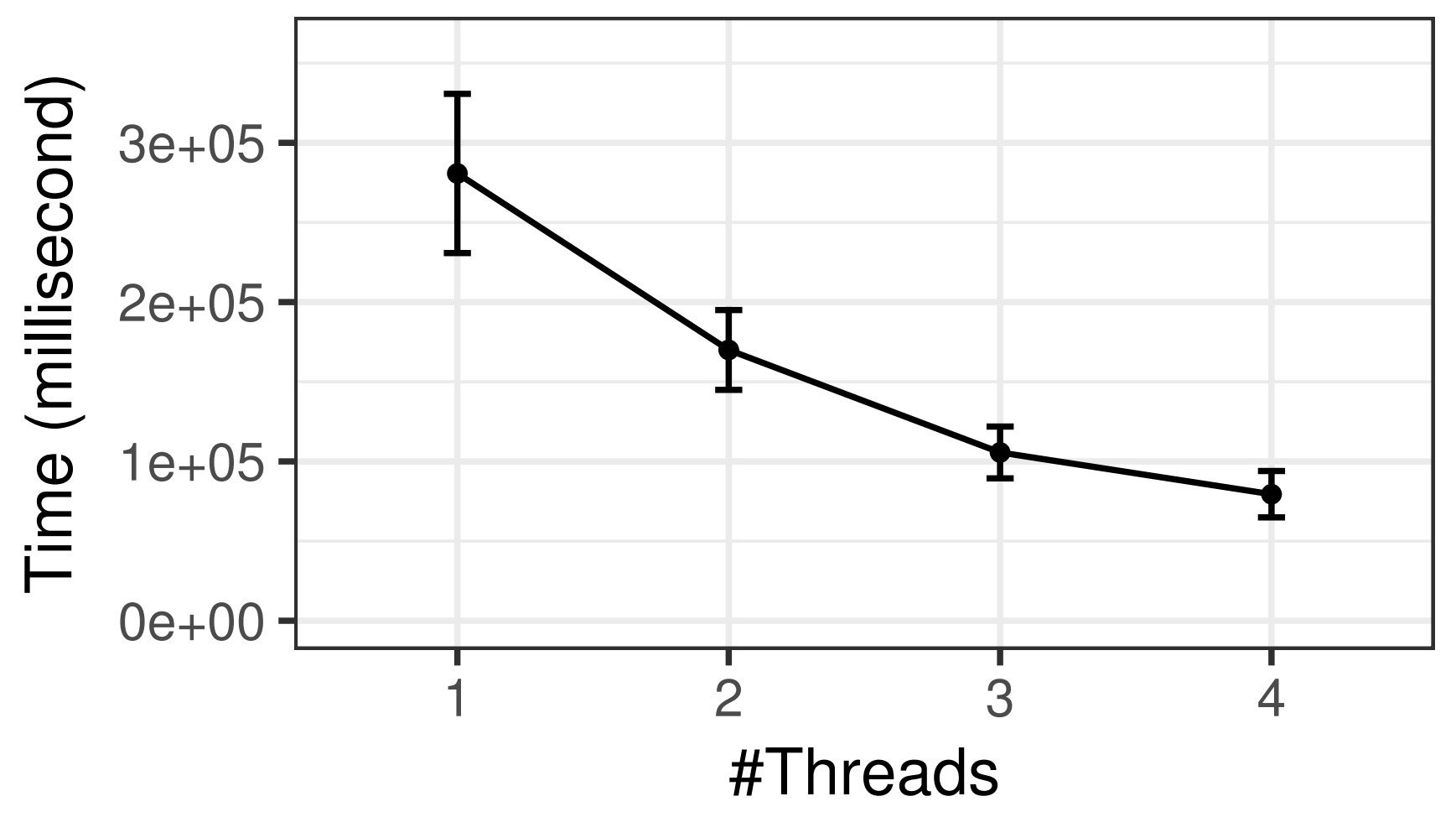
MCMC Proposals for Bayesian Phylogenetics

1. Sample particles $x_{r,k} \sim K_r(\tilde{x}_{r-1,k}, \cdot)$, a transition kernel K_r admitting π_{ϕ_r} as stationary.
2. K_r is built via a mixture of Metropolis-Hastings (MH) moves;
3. The proposals of MH moves are:
   - q_r^1: the multiplicative branch proposal;
   - q_r^2: the global multiplicative branch proposal;
   - q_r^3: the stochastic nearest neighbor interchange (NNI) proposal;
   - q_r^4: the stochastic NNI proposal with resampling the edge;
   - q_r^5: the subtree prune and regraft (SPR) move.

Experiments

- Evolutionary model: Kimura 2-parameter (K2P) model with $\theta = 2$.
- 10 taxa and 6000 sites.
- Run adaptive annealed SMC [1] with $\beta = 4$ to obtain ϕ_r $(r = 1, \cdots, R)$.

Parallelism: an advantage over MCMC



Experiments (Cont..)

Complexity analysis

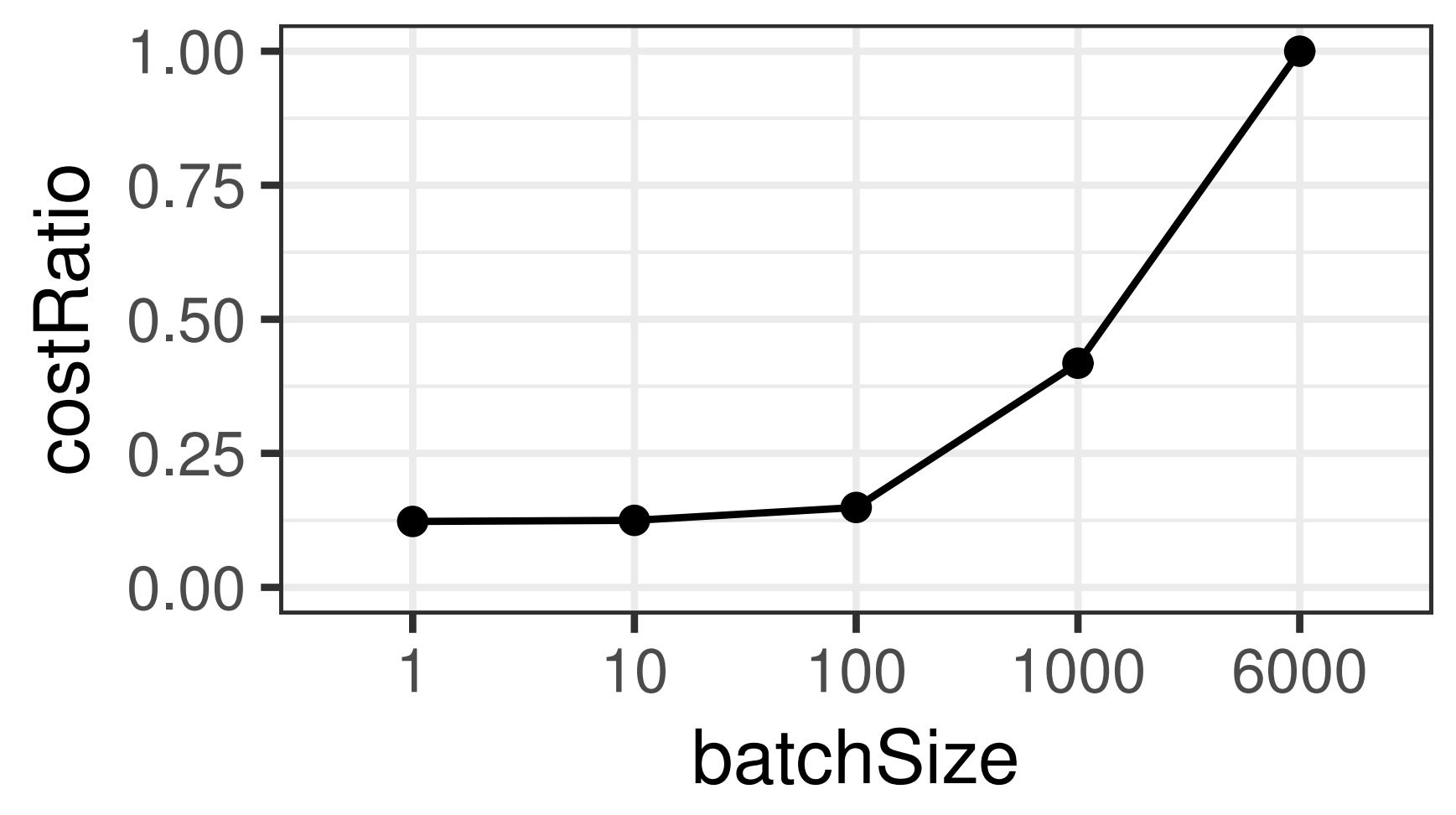


Figure 2: Ratio of cost (subsampling/annealing) as a function of batch size.

Normalizing constant & tree metric vs size of mini-batch

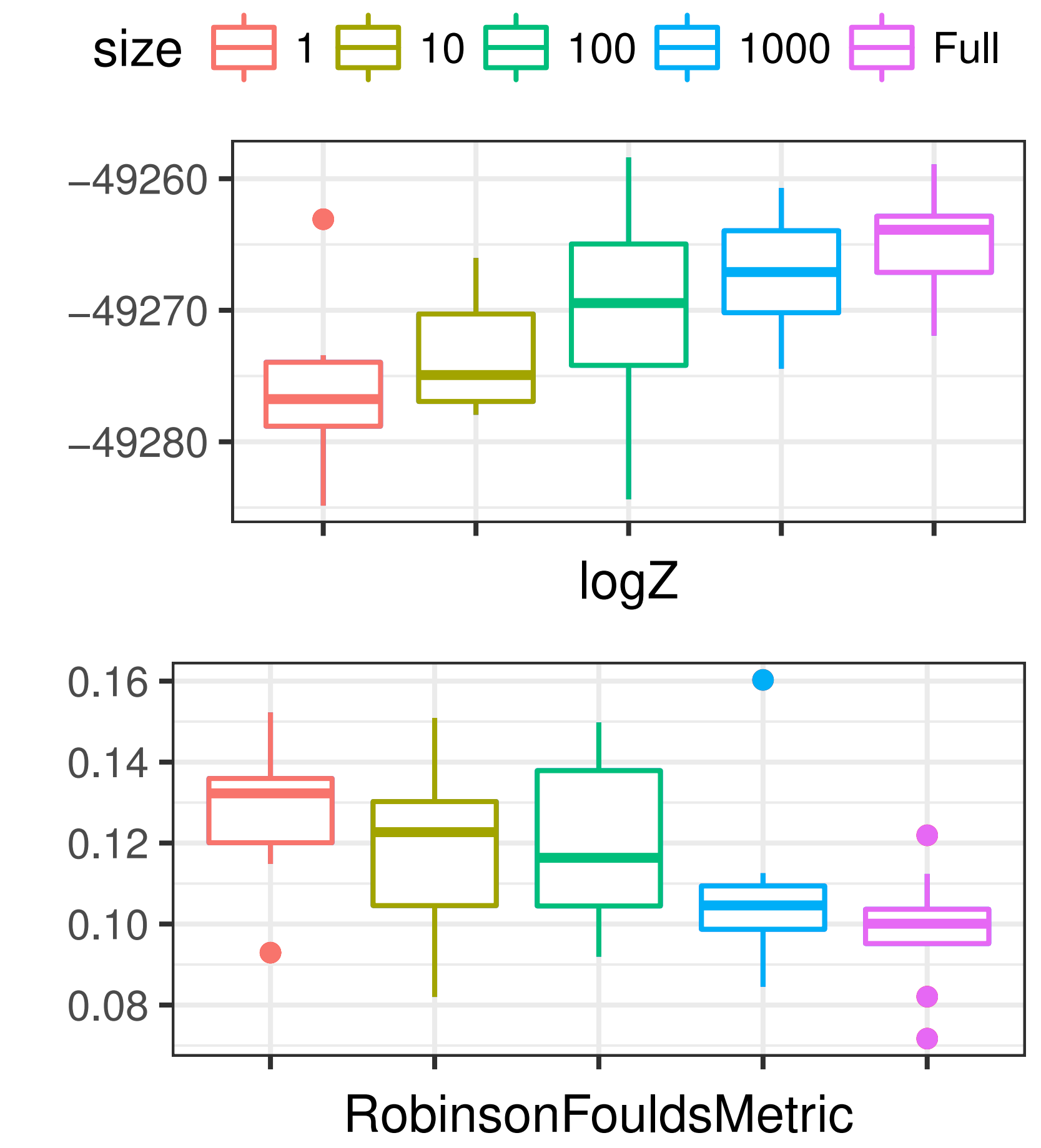


Figure 3: Comparison of subsampling SMC algorithms with different size of batch sites.

Discussion & Future work

- More scalable to phylogenetic 'tall data': parallelism; subsampling.
- To incorporate an adaptive temperature scheme ϕ_r $(r = 1, \ldots, R)$.
- Propose control variates to reduce the variability in the log-likelihood estimate.

References

[1] Liangliang Wang, Shijia Wang, and Alexandre Bouchard-Côté.
An annealed sequential Monte Carlo method for Bayesian phylogenetics.
arXiv:1806.08813, 2018.

[2] Matias Quiroz, Robert Kohn, Mattias Villani, and Minh-Ngoc Tran.
Speeding up MCMC by efficient data subsampling.
Journal of the American Statistical Association, (just-accepted):1–35, 2018.

[3] Arnaud Doucet, Nando De Freitas, and Neil Gordon.
An introduction to sequential Monte Carlo methods.
In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, 2001.

Contact Information

Shijia Wang: shijiaw@sfu.ca
http://www.sfu.ca/ shijiaw/