Variational Combinatorial Sequential Monte Carlo for Bayesian Phylogenetic Inference

Antonio Moretti, Liyi Zhang, Itsik Pe'er

Columbia University

November 23rd, 2020

Why Phylogenetic Inference?

- *Understand* how **life evolved** over time.

Why Phylogenetic Inference?

- Uncover mechanisms driving **betacoronavirus evolution**

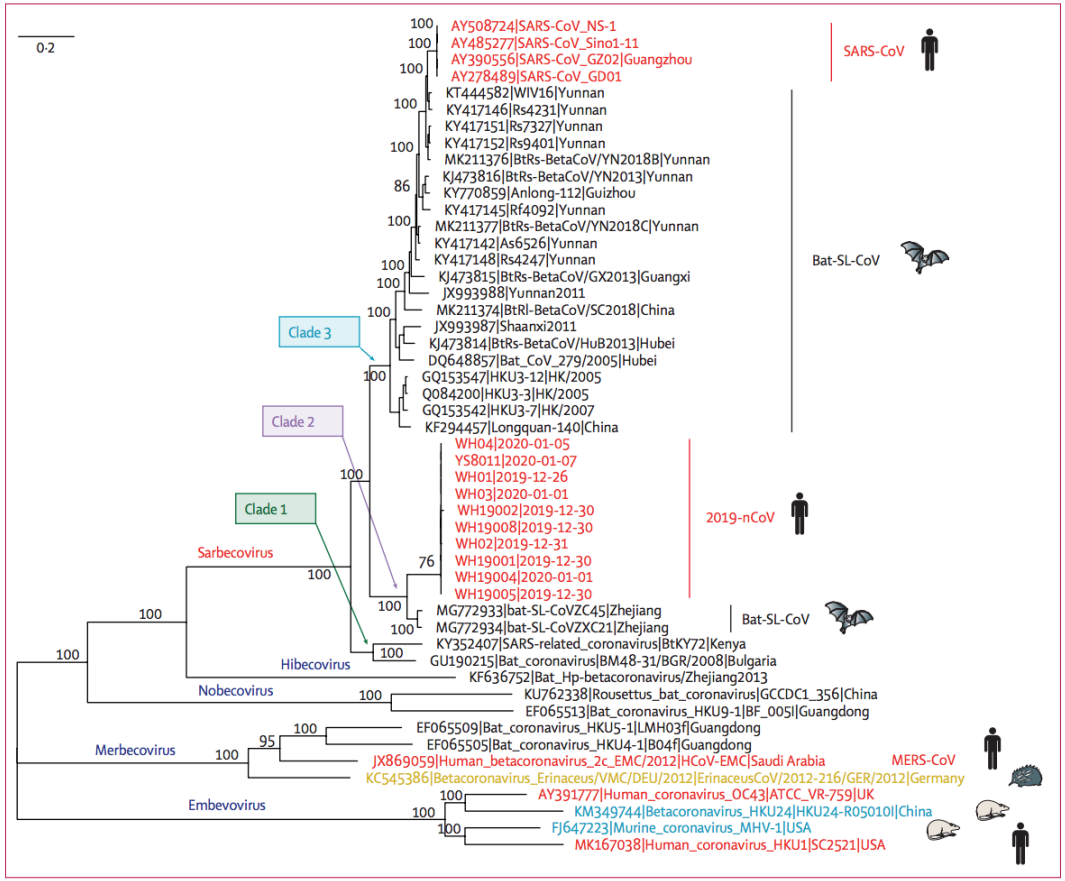
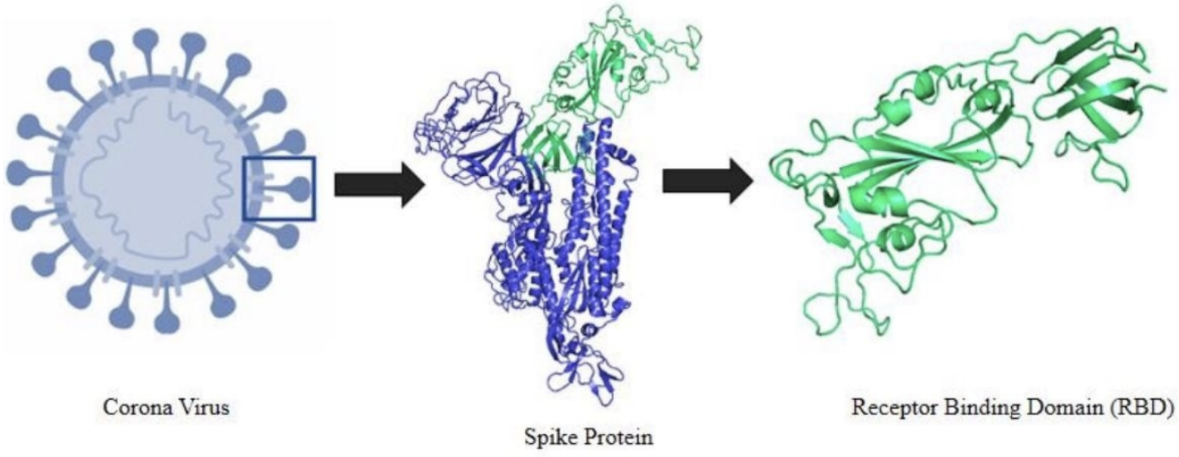


Figure 3: **Phylogenetic analysis of full-length genomes of 2019-nCoV and representative viruses of the genus Betacoronavirus**
2019-nCoV=2019 novel coronavirus. MERS-CoV=Middle East respiratory syndrome coronavirus. SARS-CoV=severe acute respiratory syndrome coronavirus.

Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, Lu et al; TheLancet, 2020. doi: 10.1016/S0140-6736(20)30251-8.

Why Phylogenetic Inference?

- Uncover mechanisms driving **betacoronavirus evolution**



- **Recombination** in RBD and **convergent evolution** $\Longrightarrow$ SARS-COV-II?

Recombination and lineage-specific mutations led to the emergence of SARS-CoV-2, Patino-Galindo et al, doi: https://doi.org/10.1101/2020.02.10.942748

Bayesian Phylogenetic Inference

- Molecular sequences $\Longrightarrow$ **evolutionary history** (DNA, RNA, PROTEIN)

$s_1 = ATGAAC$

$s_2 = ATGCAC$

$s_3 = ATGCAT$

$s_4 = ATCAAT$

Bayesian Phylogenetic Inference

- Molecular sequence data $\Longrightarrow$ **evolutionary history** (DNA, RNA, PROTEIN)

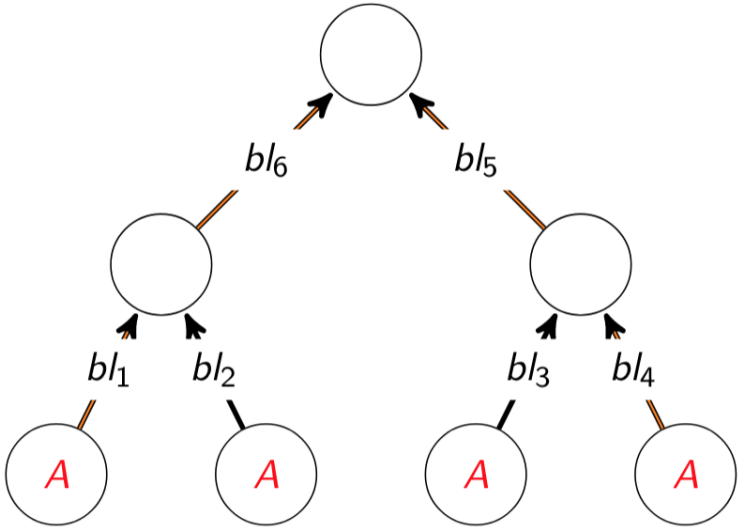
$s_1 = ATGAAC$

$s_2 = ATGCAC$

$s_3 = ATGCAT$

$s_4 = ATCAAT$

$\Longrightarrow$



- Infer **latent bifurcating tree** τ

Bayesian Phylogenetic Inference

- Molecular sequence data $\Longrightarrow$ **evolutionary history** (DNA, RNA, PROTEIN)

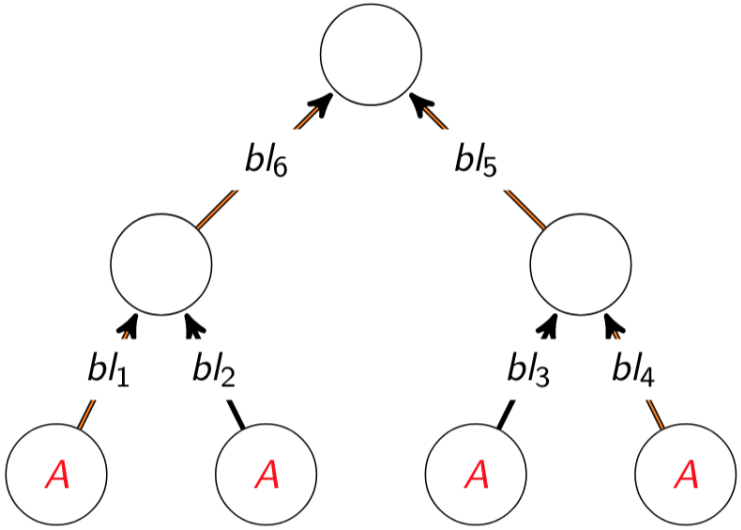
$s_1 = ATGAAC$

$s_2 = ATGCAC$

$s_3 = ATGCAT$

$s_4 = ATCAAT$

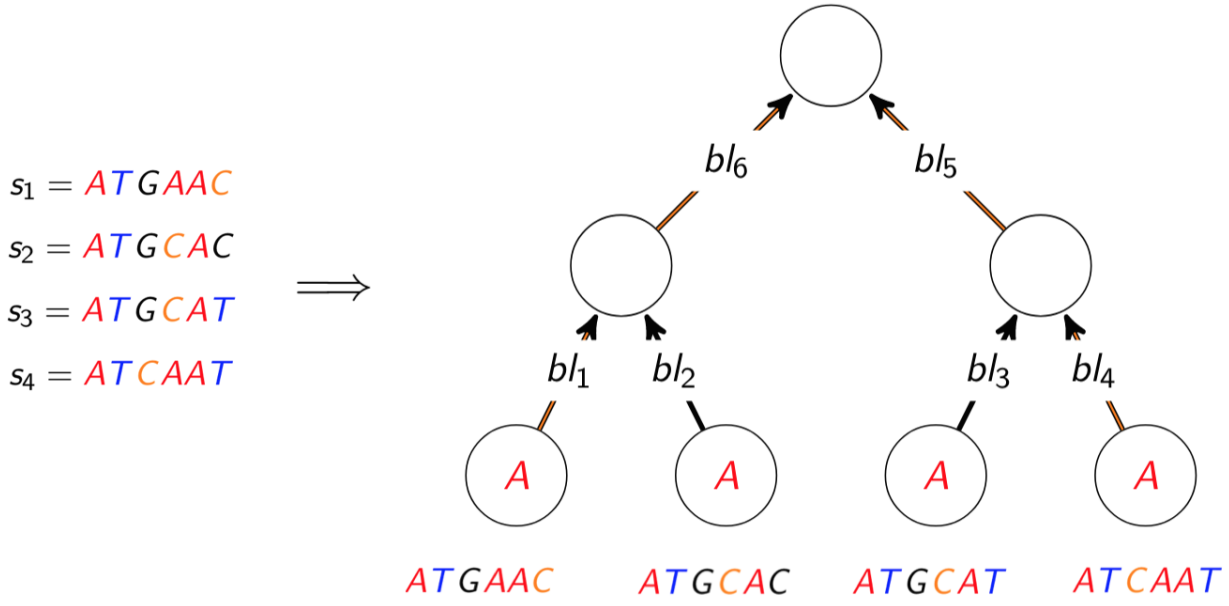
$\Longrightarrow$



- Infer latent bifurcating tree τ
  - τ a **connected acyclic graph** (V, E)

Bayesian Phylogenetic Inference

- Molecular sequence data $\Longrightarrow$ **evolutionary history** (DNA, RNA, PROTEIN)



- Infer latent bifurcating tree τ
  - τ a connected acyclic graph (V, E)
    - **Leaf nodes** are observed taxa

Bayesian Phylogenetic Inference

- Molecular sequence data $\Longrightarrow$ **evolutionary history** (DNA, RNA, PROTEIN)

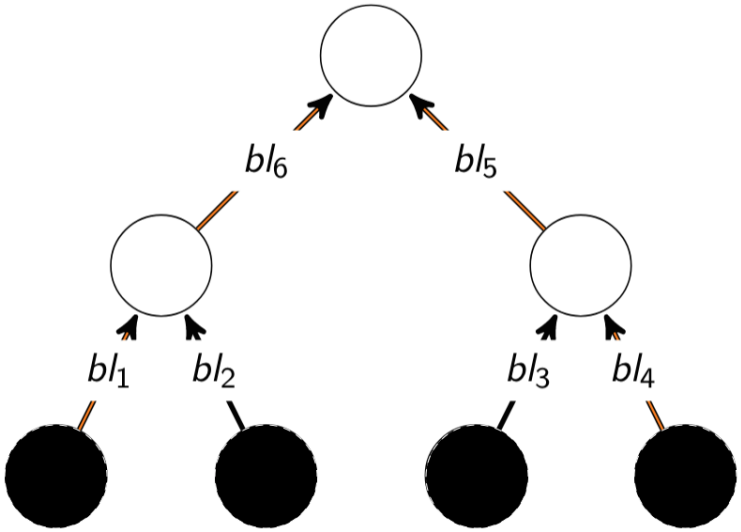
$s_1 = ATGAAC$

$s_2 = ATGCAC$

$s_3 = ATGCAT$

$s_4 = ATCAAT$

$\Longrightarrow$



- Infer latent bifurcating tree τ
  - τ a connected acyclic graph (V, E)
    - **Leaf nodes** have degree 1

Bayesian Phylogenetic Inference

- Molecular sequence data $\Longrightarrow$ **evolutionary history** (DNA, RNA, PROTEIN)

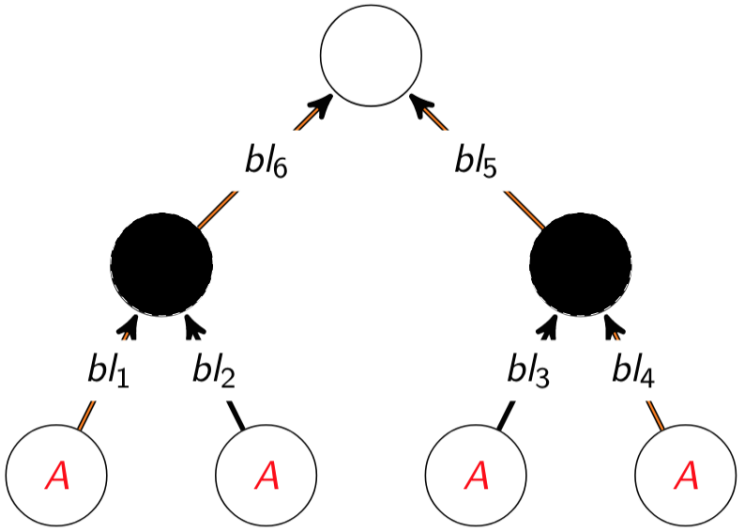
$s_1 = ATGAAC$

$s_2 = ATGCAC$

$s_3 = ATGCAT$

$s_4 = ATCAAT$

$\Longrightarrow$



- Infer latent bifurcating tree τ
  - τ a connected acyclic graph (V, E)
    - **Internal nodes** are *unobserved ancestral taxa*

Bayesian Phylogenetic Inference

- Molecular sequence data $\Longrightarrow$ **evolutionary history** (DNA, RNA, PROTEIN)

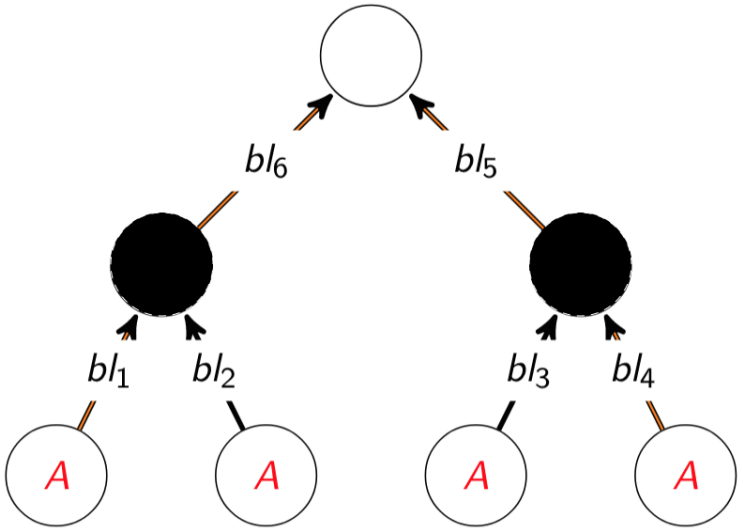
$s_1 = ATGAAC$

$s_2 = ATGCAC$

$s_3 = ATGCAT$

$s_4 = ATCAAT$

$\Longrightarrow$



- Infer latent bifurcating tree τ
  - τ a connected acyclic graph (V, E)
    - **Internal nodes** have degree 3

Bayesian Phylogenetic Inference

- Molecular sequence data $\Longrightarrow$ **evolutionary history** (DNA, RNA, PROTEIN)

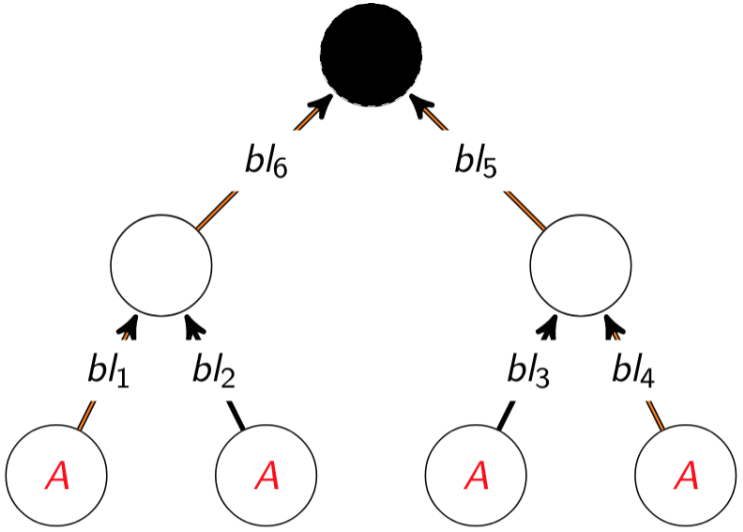
$s_1 = ATGAAC$

$s_2 = ATGCAC$

$s_3 = ATGCAT$

$s_4 = ATCAAT$

$\Longrightarrow$



- Infer latent bifurcating tree τ
  - τ a connected acyclic graph (V, E)
    - **Root node** is *common evolutionary ancestor*

Bayesian Phylogenetic Inference

- Molecular sequence data $\Longrightarrow$ **evolutionary history** (DNA, RNA, PROTEIN)

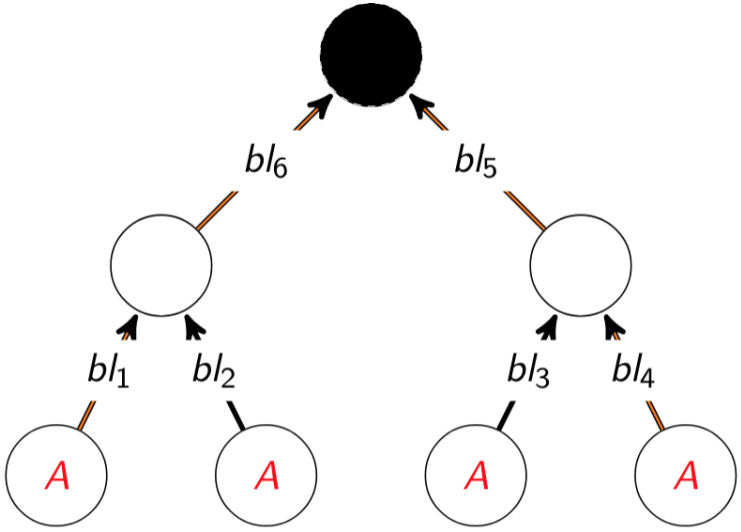
$s_1 = ATGAAC$

$s_2 = ATGCAC$

$s_3 = ATGCAT$

$s_4 = ATCAAT$

$\Longrightarrow$



- Infer latent bifurcating tree τ
  - τ a connected acyclic graph (V, E)
    - **Root node** has degree 2

Bayesian Phylogenetic Inference

- Molecular sequence data $\Longrightarrow$ **evolutionary history** (DNA, RNA, PROTEIN)

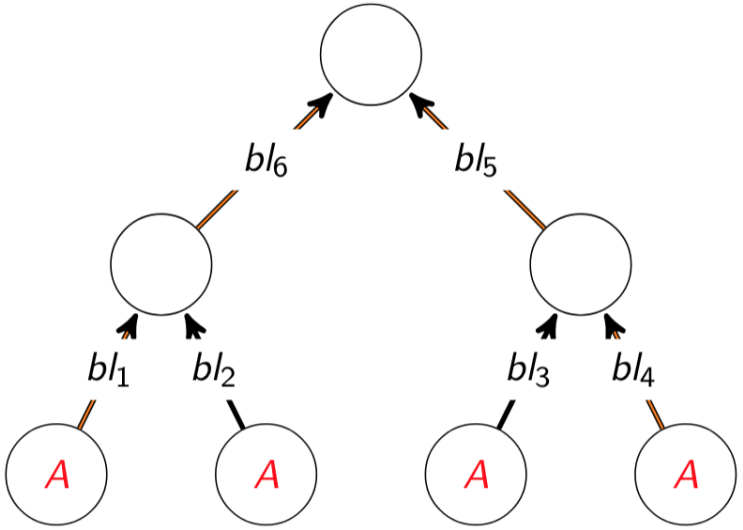
$s_1 = ATGAAC$

$s_2 = ATGCAC$

$s_3 = ATGCAT$

$s_4 = ATCAAT$

$\Longrightarrow$



- Infer latent bifurcating tree τ
  - τ a connected acyclic graph (V, E)
  - $|E|$ **branch lengths** $b(e) \in \mathbb{R}_{>0}, b(e) \in \mathcal{B}$

Bayesian Phylogenetic Inference

- Molecular sequence data $\Longrightarrow$ **evolutionary history** (DNA, RNA, PROTEIN)

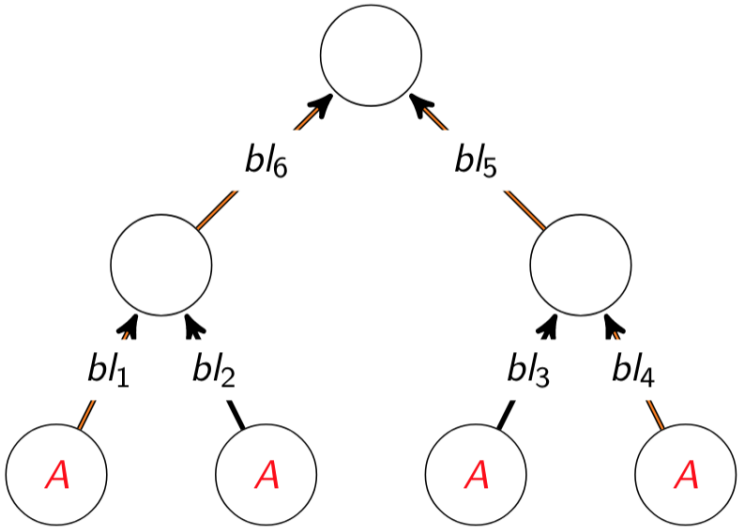
$s_1 = ATGAAC$

$s_2 = ATGCAC$

$s_3 = ATGCAT$

$s_4 = ATCAAT$

$\Longrightarrow$



- Infer latent bifurcating tree τ
  - τ a connected acyclic graph (V, E)
  - **Nonclock trees** have *nonconstant evolutionary rate*

Evolutionary Model

- Given a tree τ on data $\mathrm{Y} = \{Y_1, \cdots, Y_M\} \in \Omega^{NxM}$

Evolutionary Model

- Given a tree τ on data $\mathrm{Y} = \{Y_1, \cdots, Y_M\} \in \Omega^{N x M}$

$\Longrightarrow$ Need **model** to specify **data likelihood**:

$$p(\mathrm{Y}|\tau, \mathcal{B}, \theta) = \prod_{i=1}^{M} p(Y_i|\tau, \mathcal{B}, \theta)$$

Evolutionary Model

- Given a tree τ on data $\mathrm{Y} = \{Y_1, \cdots, Y_M\} \in \Omega^{NxM}$

$\Longrightarrow$ Need **model** to specify **data likelihood**:

$$p(\mathrm{Y}|\tau, \mathcal{B}, \theta) = \prod_{i=1}^{M} p(Y_i|\tau, \mathcal{B}, \theta)$$

$\Longrightarrow$ Define **prob of transition** b/t characters (*nucleotides*):

  - CTMC with rate matrix Q

Evolutionary Model

- Given a tree τ on data $\mathrm{Y} = \{Y_1, \cdots, Y_M\} \in \Omega^{NxM}$

$\Longrightarrow$ Need **model** to specify **data likelihood**:

$$p(\mathrm{Y}|\tau, \mathcal{B}, \theta) = \prod_{i=1}^{M} p(Y_i|\tau, \mathcal{B}, \theta)$$

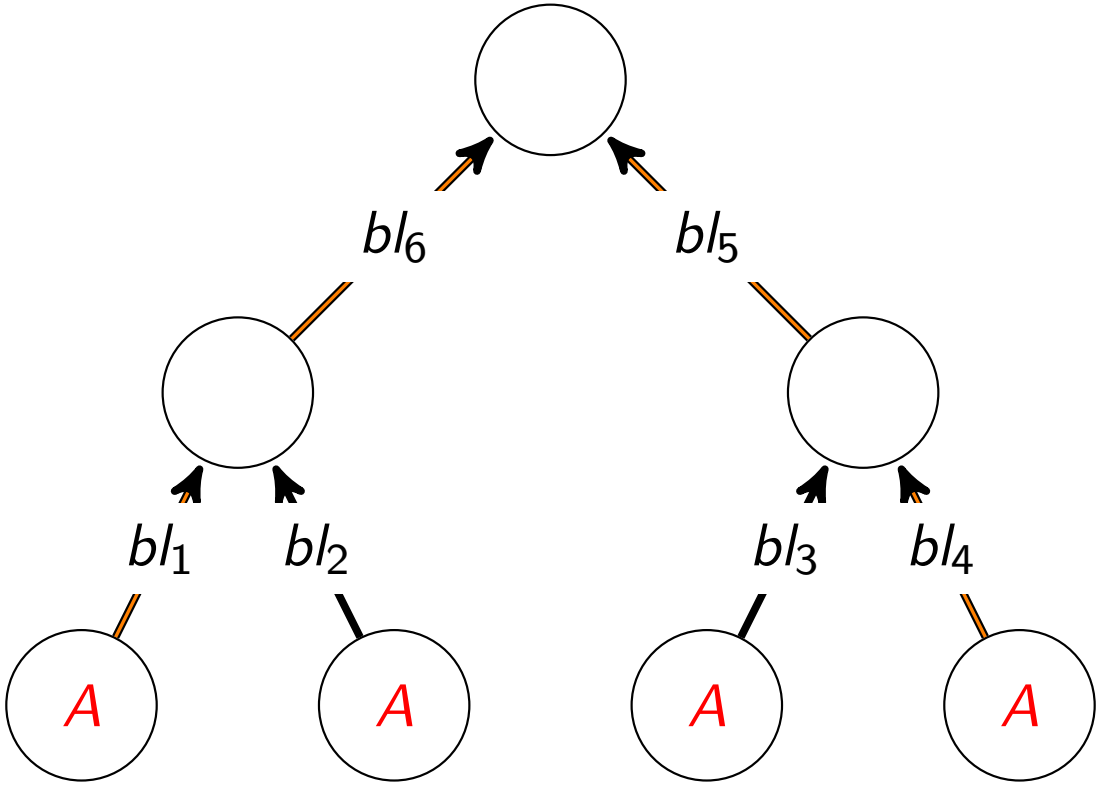
$\Longrightarrow$ Define **prob of transition** b/t characters (*nucleotides*):

- CTMC with rate matrix Q

Let $\zeta_{v,s}$ be state of genome for species v at site s:

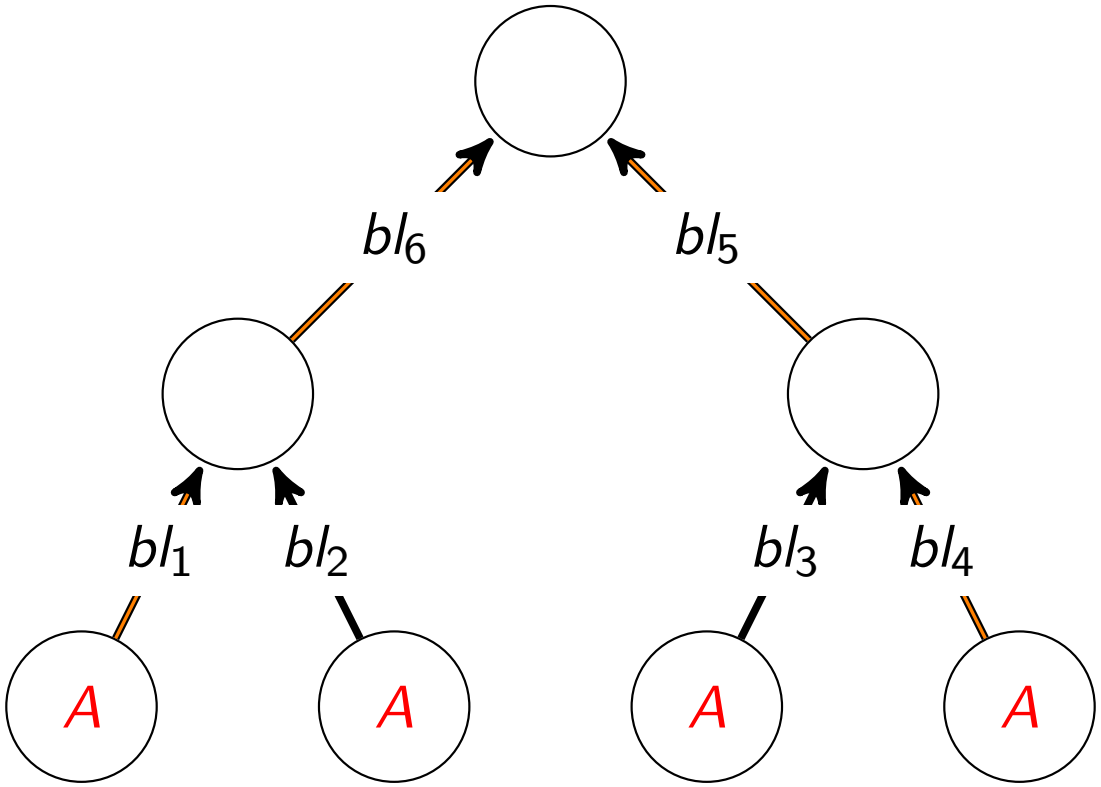
$$P(\zeta_{v',s} = j|\zeta_{v,s} = i) = (exp\,(b(e)Q))_{i,j}$$

Computing the Likelihood



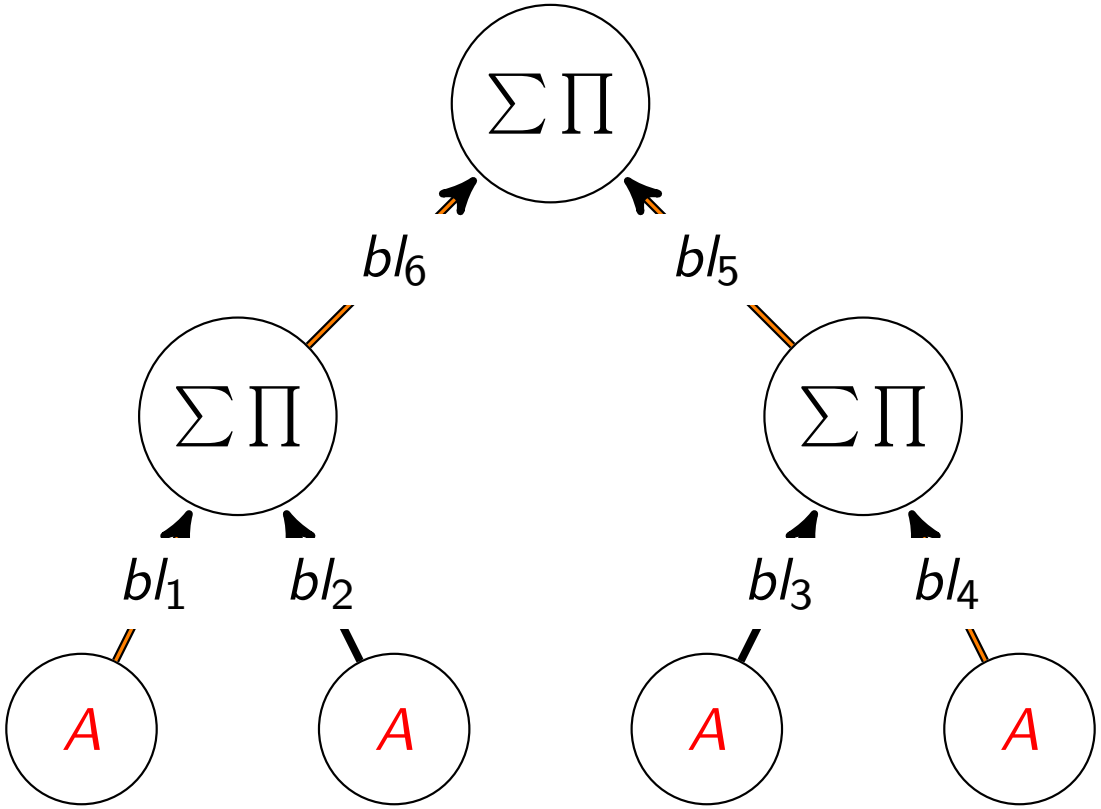
$$P(\mathrm{Y}|\tau, \mathcal{B}, \theta) := \prod_{i=1}^{M} \sum_{a^i} \eta(a^i_\rho) \prod_{(u,v)\in E(\tau)} \exp\left(-b_{u,v} \mathrm{Q}_{a^i_u, a^i_v}\right)$$

Computing the Likelihood



- Sum-Product / Belief Propagation / Pruning Algorithm

Computing the Likelihood



- Pass messages for conditional likelihood at site i:

$$L_P(i) = \left(\sum_{x \in k} Pr(x|i, t_L) L_L(x))\right) \cdot \left(\sum_{x \in k} Pr(x|i, t_R) L_R(x))\right)$$

The Bayesian Approach

- How many **distinct tree topologies**?

The Bayesian Approach

- How many **distinct tree topologies**?

  $(2N - 3)!!$

The Bayesian Approach

- How many **distinct tree topologies**?

  $(2N-3)!!$

- **Evolutionary uncertainty** and prior information

$$p(\mathcal{B}, \tau, \theta | \mathrm{Y}) = \frac{p(\mathrm{Y}|\tau, \mathcal{B}, \theta) p(\tau, \mathcal{B}|\theta) p(\theta)}{p(\mathrm{Y})}$$

The Bayesian Approach

- How many **distinct tree topologies**?

  $(2N-3)!!$

- Posterior over phylogenies:

$$p(\mathcal{B},\tau,\theta|\mathrm{Y}) = \frac{\overbrace{p(\mathrm{Y}|\tau,\mathcal{B},\theta)}^{\textit{Likelihood}}\ \overbrace{p(\tau,\mathcal{B}|\theta)p(\theta)}^{\textit{tree \& model prior}}}{\underbrace{p(\mathrm{Y})}_{\textit{evidence}}}$$

The Bayesian Approach

- How many **distinct tree topologies**?

  $(2N-3)!!$

- Posterior over phylogenies:

$$p(\mathcal{B}, \tau, \theta | \mathrm{Y}) = \frac{p(\mathrm{Y}|\tau, \mathcal{B}, \theta) p(\tau, \mathcal{B}|\theta) p(\theta)}{p(\mathrm{Y})}$$

- **Marginalizing** $p(\mathrm{Y})$ intractable.

$$P(Y) = \sum_{\tau \in \mathcal{T}} \int p(\mathrm{Y}|\tau, \mathcal{B}, \theta) p(\tau, \mathcal{B}|\theta) p(\theta) d\theta d\tau$$

Bayesian Phylogenetic Inference

Several distinct challenges:

Bayesian Phylogenetic Inference

Several distinct challenges:

- **Inference** (*marginalization*)

Bayesian Phylogenetic Inference

Several distinct challenges:

- **Inference** (*marginalization*)
  - Sample to approx sum over **tree topologies** τ

Bayesian Phylogenetic Inference

Several distinct challenges:

- **Inference** (*marginalization*)
  - Sample to approx sum over **tree topologies** τ
  - For each τ, sample to approx integral over **branch lengths**

Bayesian Phylogenetic Inference

Several distinct challenges:

- **Inference** (*marginalization*)
  - Sample to approx sum over **tree topologies** τ
  - For each τ, sample to approx integral over **branch lengths**
- **Learning** (*optimization*)

Bayesian Phylogenetic Inference

Several distinct challenges:

- **Inference** (*marginalization*)
  - Sample to approx sum over **tree topologies** τ
  - For each τ, sample to approx integral over **branch lengths**

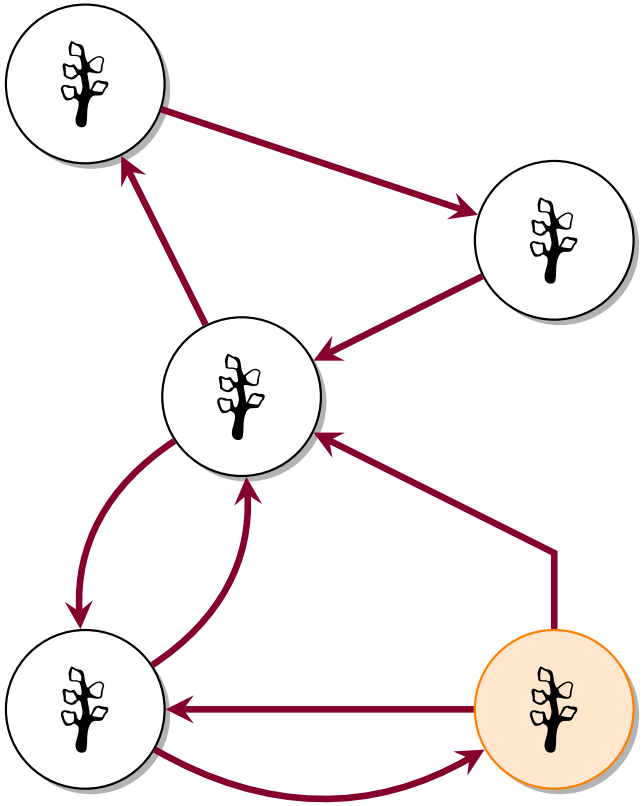
- **Learning** (*optimization*)
  - Find parameters $\theta = (Q, \{\lambda_i\}_{i=1}^{|E|} \in \mathcal{B})$ to **max data likelihood**

Approaches: Local vs Sequential Search

- **Local search**: MCMC

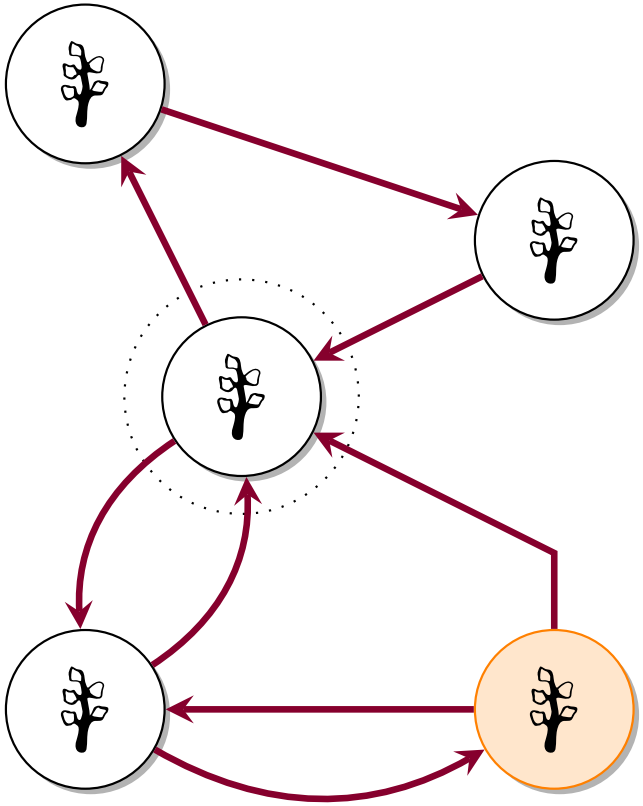
Approaches: Local vs Sequential Search

- **Local search**: MCMC
  - Start w/ initial τ



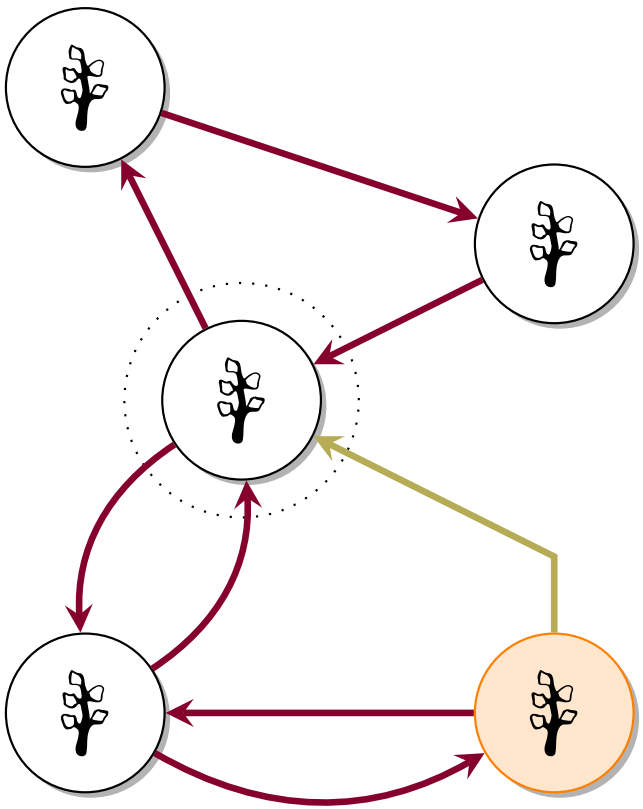
Approaches: Local vs Sequential Search

- **Local search**: MCMC
  - Sample $\tau' \sim q(\cdot|\tau^i)$ by perturbing τ^i



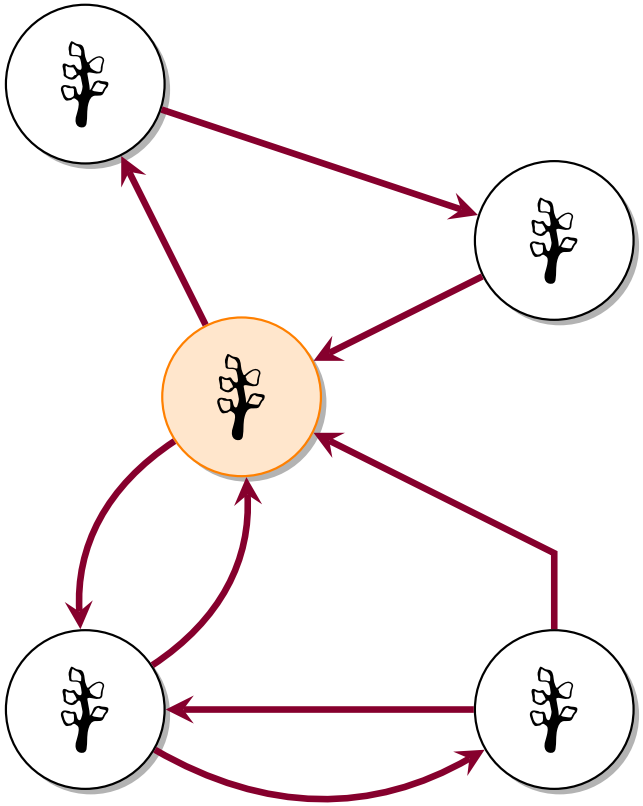
Approaches: Local vs Sequential Search

- **Local search**: MCMC
  - Sim $U \sim \text{UNIFORM}(0, 1)$
    move to τ' if $U \leq \alpha(\tau', \tau^i)$



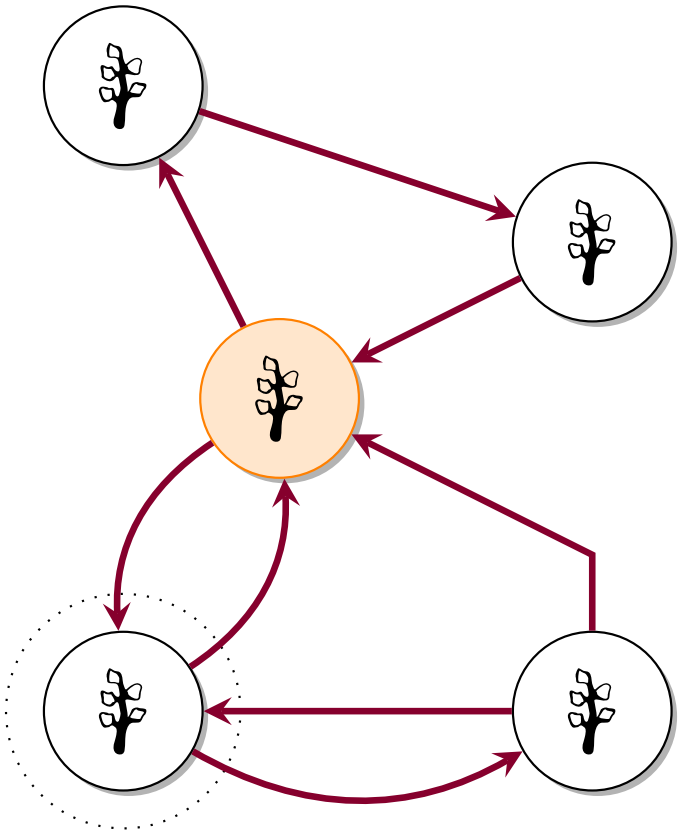
Approaches: Local vs Sequential Search

- **Local search**: MCMC
  - If accept, set $\tau^{i+1} = \tau'$



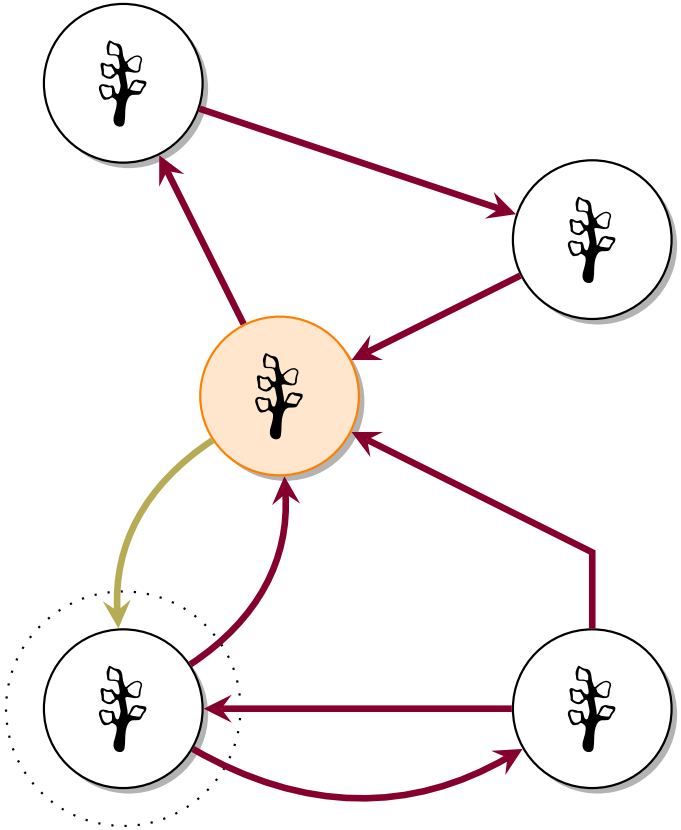
Approaches: Local vs Sequential Search

- **Local search**: MCMC
  - Sample $\tau' \sim q(\cdot|\tau^i)$ by perturbing τ^i



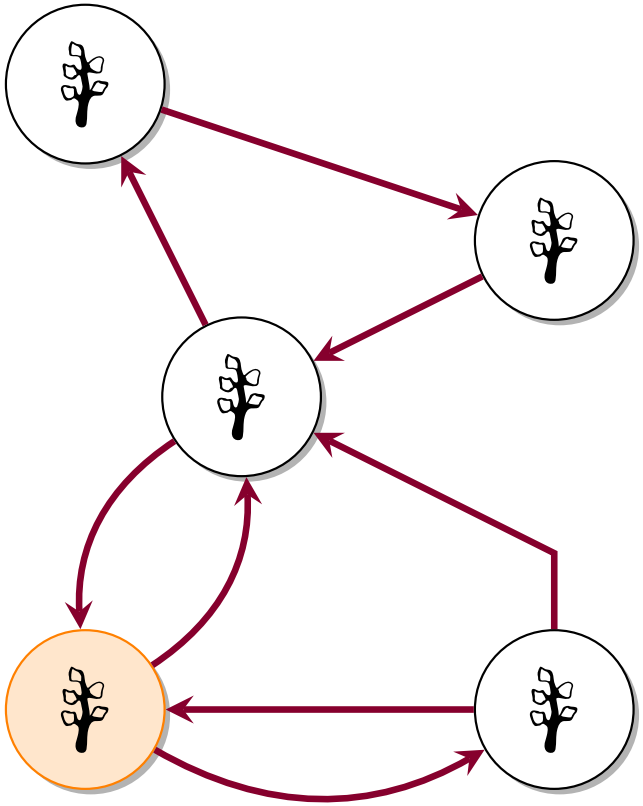
Approaches: Local vs Sequential Search

- **Local search**: $\textsc{Mcmc}$
  - Sim $U \sim \textsc{Uniform}(0, 1)$
    move to τ' if $U \leq \alpha(\tau', \tau^i)$



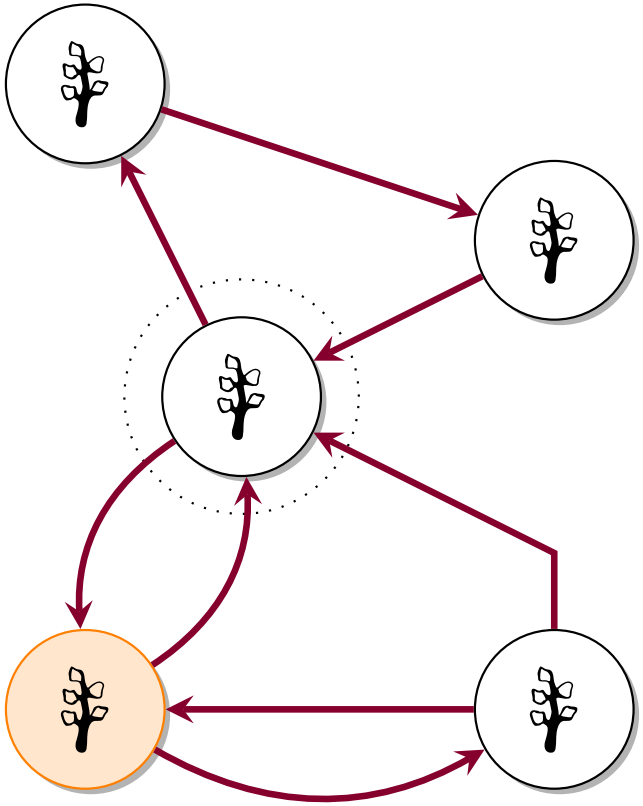
Approaches: Local vs Sequential Search

- **Local search**: MCMC
  - If accept, set $\tau^{i+1} = \tau'$



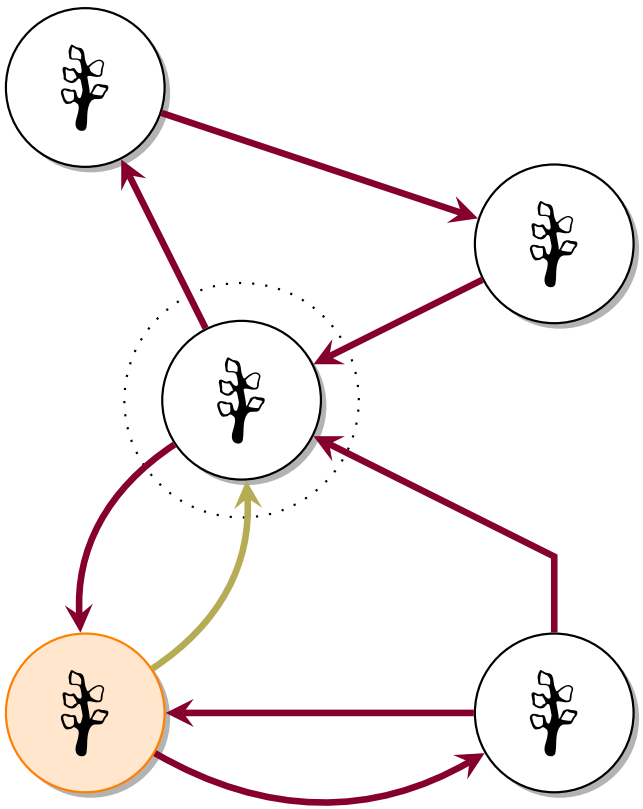
Approaches: Local vs Sequential Search

- **Local search**: MCMC
  - Sample $\tau' \sim q(\cdot|\tau^i)$ by perturbing τ^i



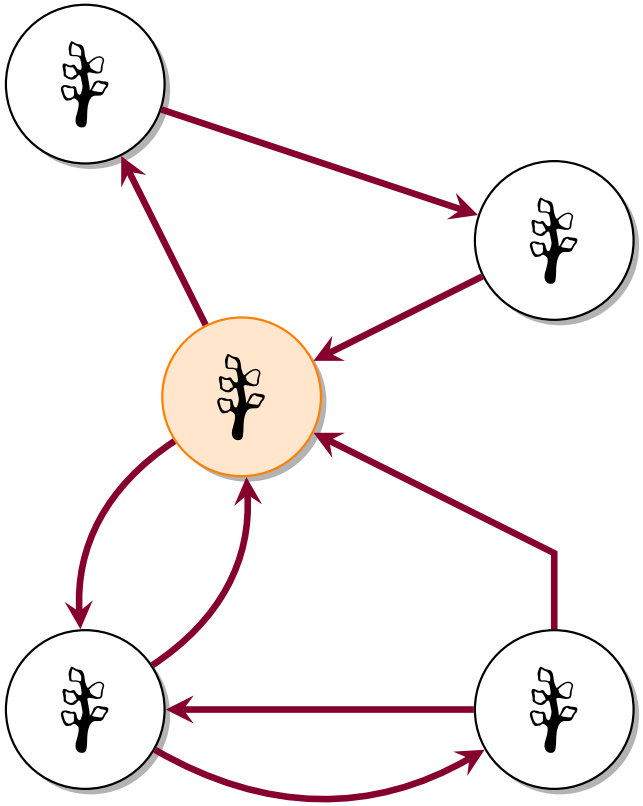
Approaches: Local vs Sequential Search

- **Local search**: $\textsc{Mcmc}$
  - Sim $U \sim \textsc{Uniform}(0, 1)$
    move to τ' if $U \leq \alpha(\tau', \tau^i)$



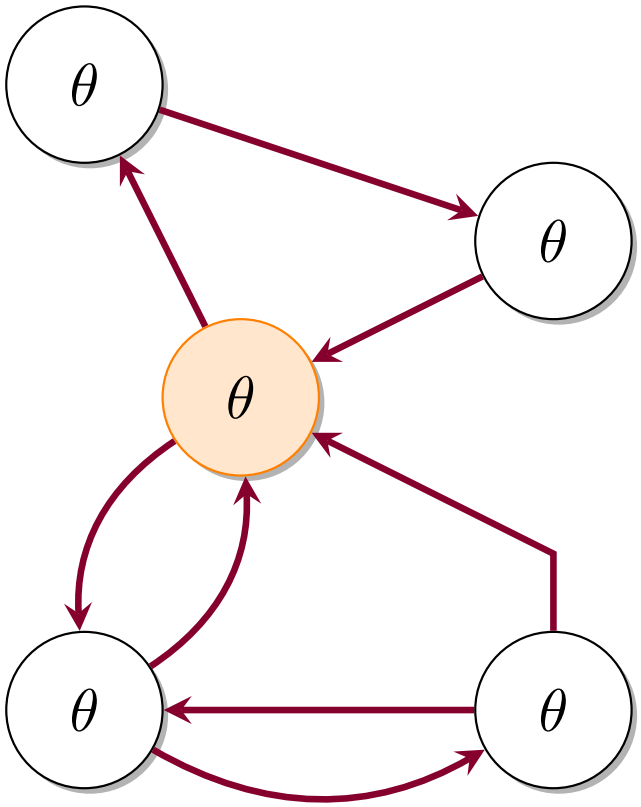
Approaches: Local vs Sequential Search

- **Local search**: MCMC
  - If accept, set $\tau^{i+1} = \tau'$



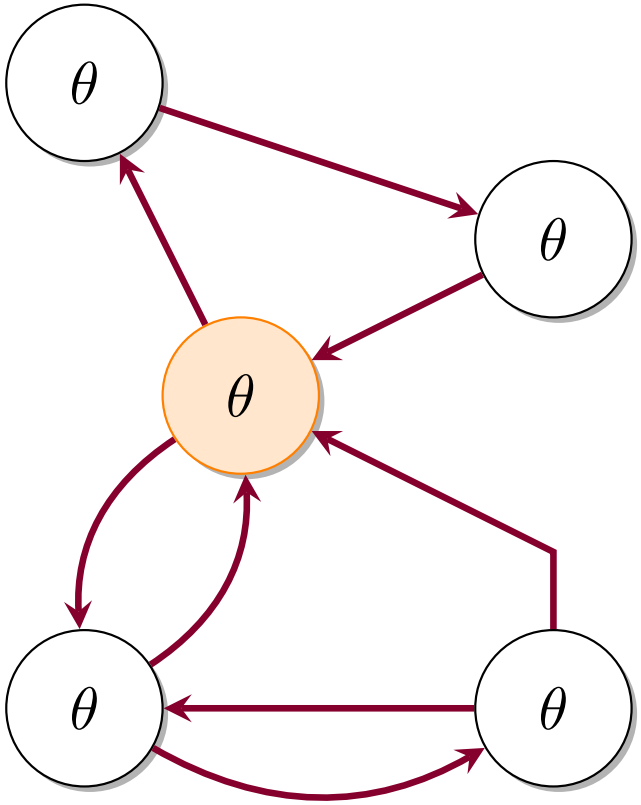
Approaches: Local vs Sequential Search

- **Local search**: MCMC
  - Can be used for both *inference* and *learning*



Approaches: Local vs Sequential Search

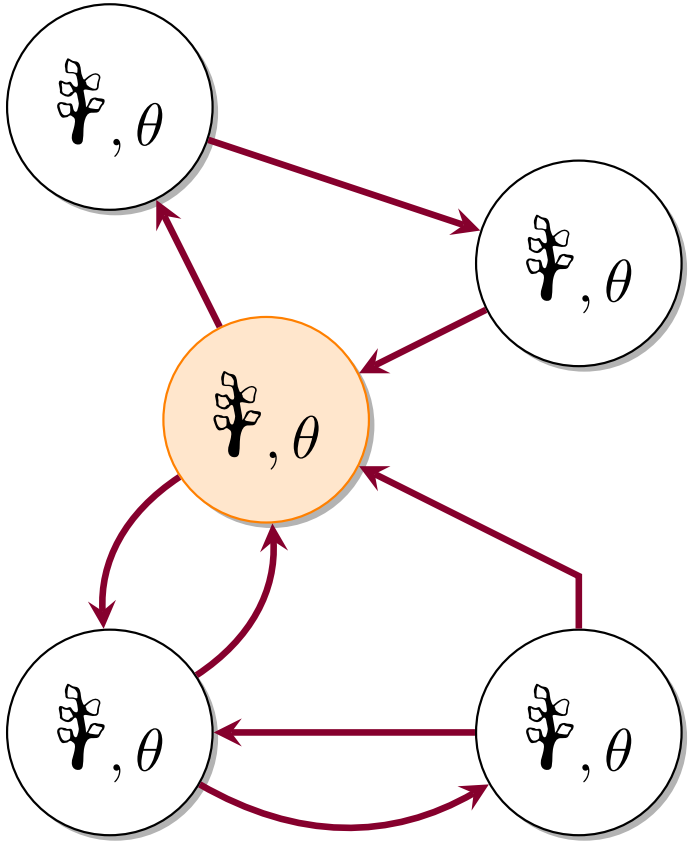
- **Local search**: MCMC
  - Can be used for both *inference* and *learning*



- *Long runs* and **inefficient parameter space exploration**

Approaches: Local vs Sequential Search

- **Local search**: MCMC
  - Can be used for both *inference* and *learning*



$\Longrightarrow$ **Complex**, **multimodal** dist on *composite space*.

Approaches: Local vs Sequential Search

- **Local search**: MCMC
  - Can be used for both *inference* and *learning*
    - Mr Bayes (Huelsenbeck & Ronquist, 2001)

Approaches: Local vs Sequential Search

- MCMC is **local search** algorithm
  - Can be used for both *inference* and *learning*
    - Mr Bayes (Huelsenbeck & Ronquist, 2001)
    - Probabilistic Path Hamiltonian Monte Carlo (Dinh et al., 2017)

Approaches: Local vs Sequential Search

- **Local search**: MCMC
  - Can be used for both *inference* and *learning*
    - Mr Bayes (Huelsenbeck & Ronquist, 2001)
    - Probabilistic Path Hamiltonian Monte Carlo (Dinh et al., 2017)

- **Sequential search**: SMC

Approaches: Local vs Sequential Search

- **Local search**: MCMC
  - Can be used for both inference and learning
    - Mr Bayes (Huelsenbeck & Ronquist, 2001)
    - Probabilistic Path Hamiltonian Monte Carlo (Dinh et al., 2017)

- **Sequential search**: SMC
  - Performs *inference* but *requires* MCMC or EM step for learning.

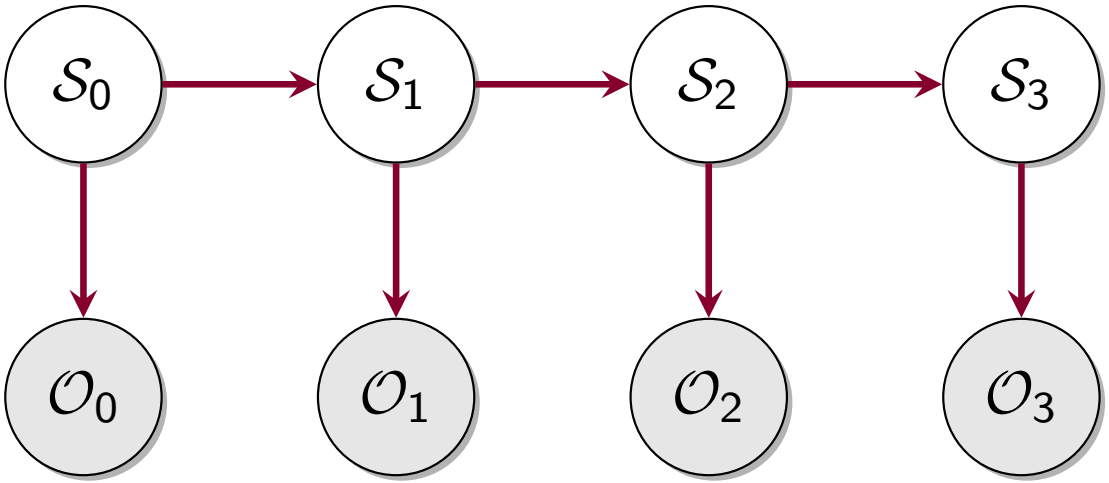
Approaches: Local vs Sequential Search

- **Local search:** MCMC
  - Can be used for both *inference* and *learning*
    - Mr Bayes (Huelsenbeck & Ronquist, 2001)
    - Probabilistic Path Hamiltonian Monte Carlo (Dinh et al., 2017)

- **Sequential search**: SMC
  - Performs *inference* but requires MCMC or EM step for learning
    - Poset SMC (Bouchard-Cote, 2012)
    - Combinatorial SMC (Wang, 2015)
  - Particle MCMC approaches
    - ⇒ Use SMC for *inference* & MCMC for *learning*.
    - CSMC (Wang, 2015), Particle Gibbs (Wang, 2020)

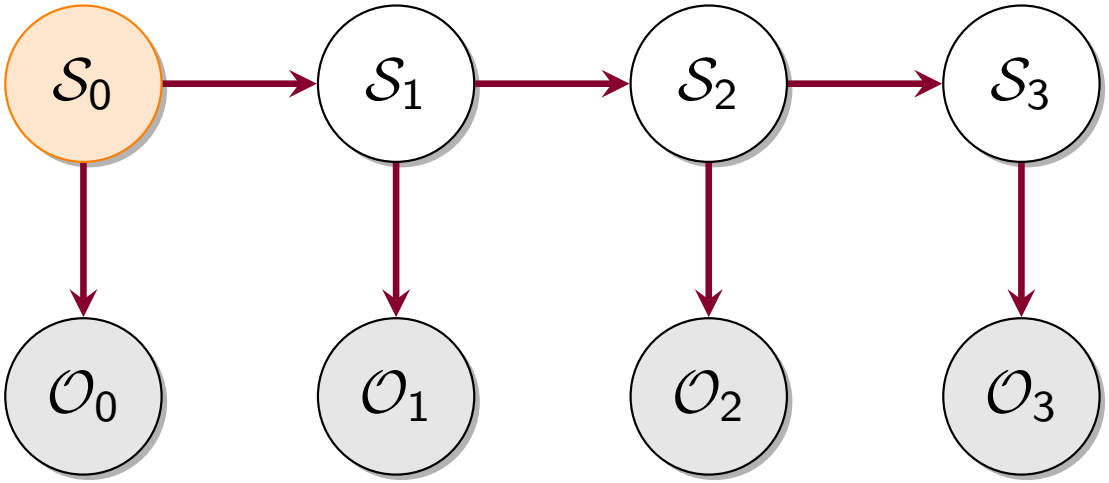
Sequential Search: Combinatorial SMC

- SMC operates on a **sequence of probability spaces**



Sequential Search: Combinatorial SMC

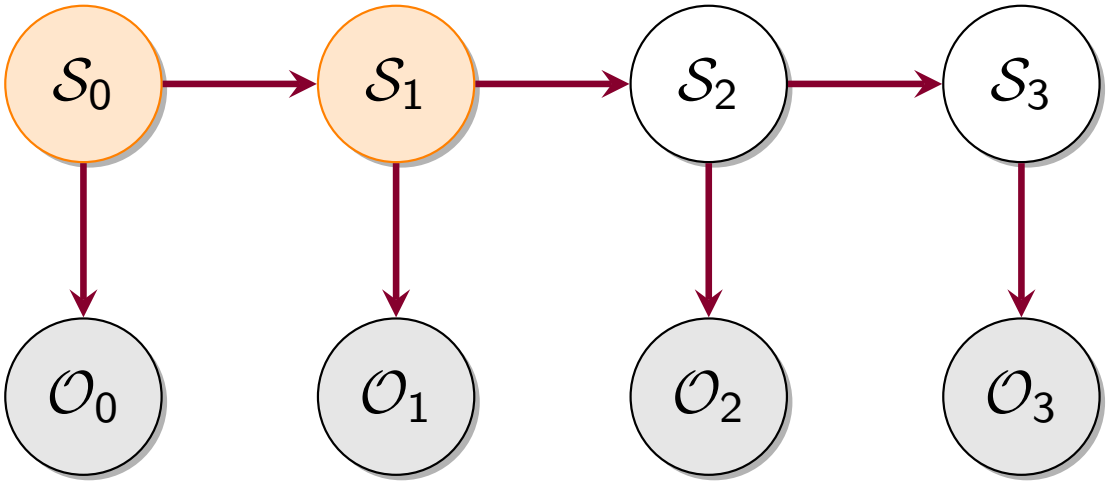
- Decompose **phylogeny space** $\mathcal{X}$ into set of **partial states** of rank r denoted $\mathcal{S}_r$, $w/\ \mathcal{S} = \bigcup_{r=1}^{R} \mathcal{S}_r$

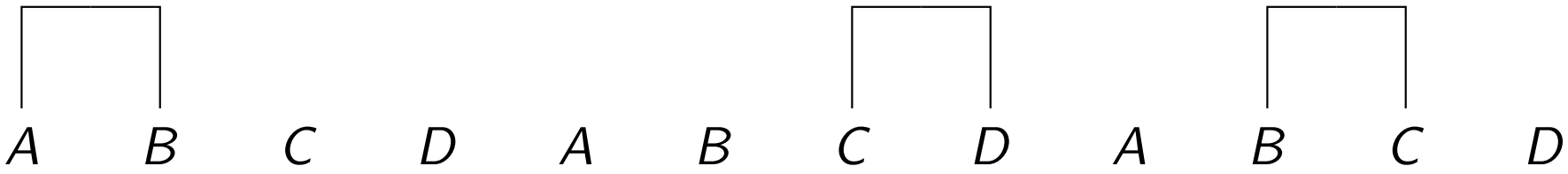


$A \quad B \quad C \quad D \quad A \quad B \quad C \quad D \quad A \quad B \quad C \quad D$

Sequential Search: Combinatorial SMC

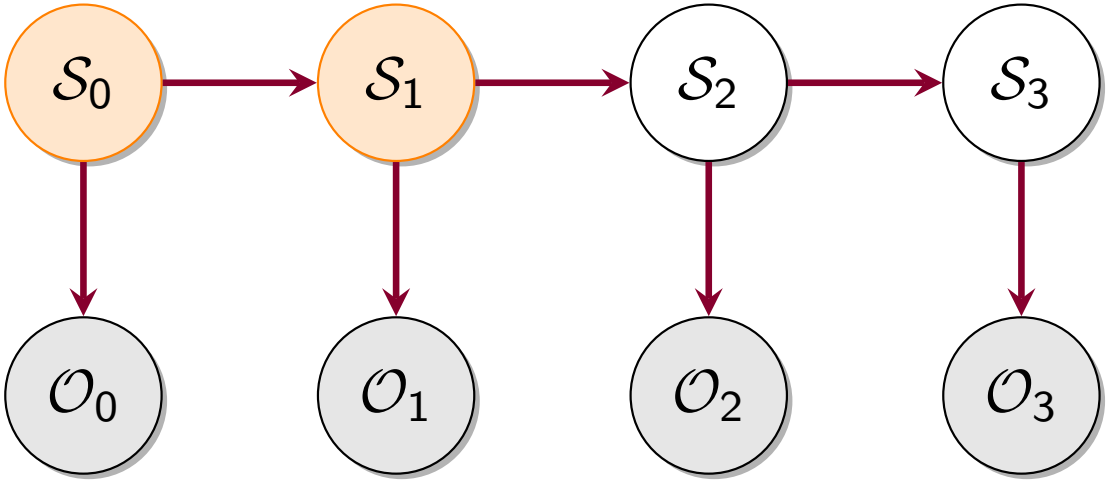
- Draw K **partial states** $\{s_{r,k}\}_{k=1}^{K} \in \mathcal{S}_r$ at each rank r

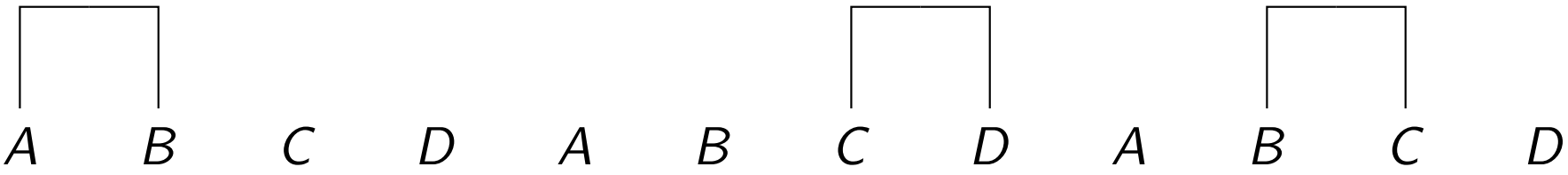




Sequential Search: Combinatorial SMC

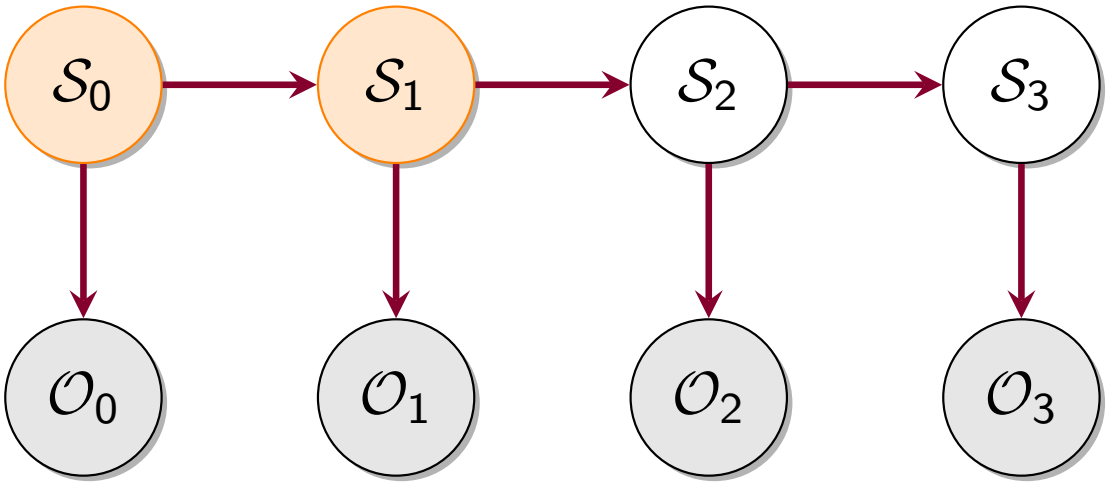
- Assign **importance weight** $\{w_{r,k}\}_{k=1}^{K}$ to each **partial state** $\{s_{r,k}\}_{k=1}^{K} \in \mathcal{S}_r$

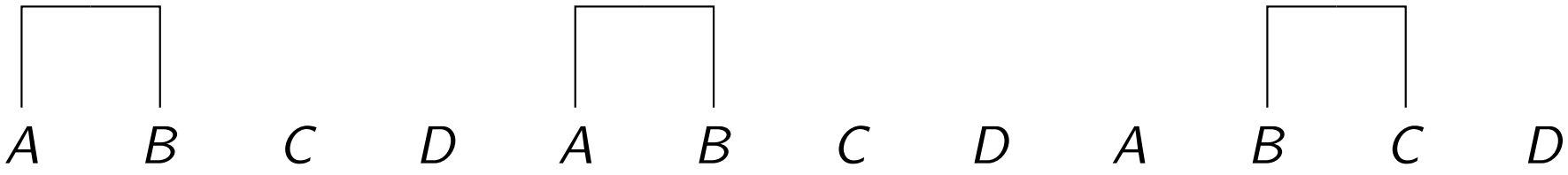




Sequential Search: Combinatorial SMC

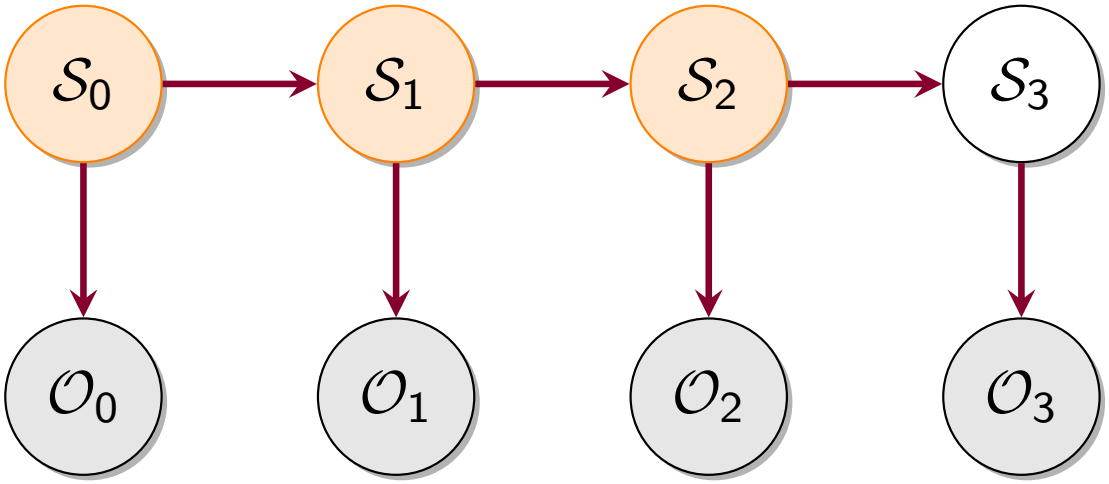
- **Resample** state $\tilde{s}_{r,k} \sim \textsc{Categorical}(\bar{w}_{r-1,1}, \cdots, \bar{w}_{r-1,K})$ to focus on areas of **high probability**.

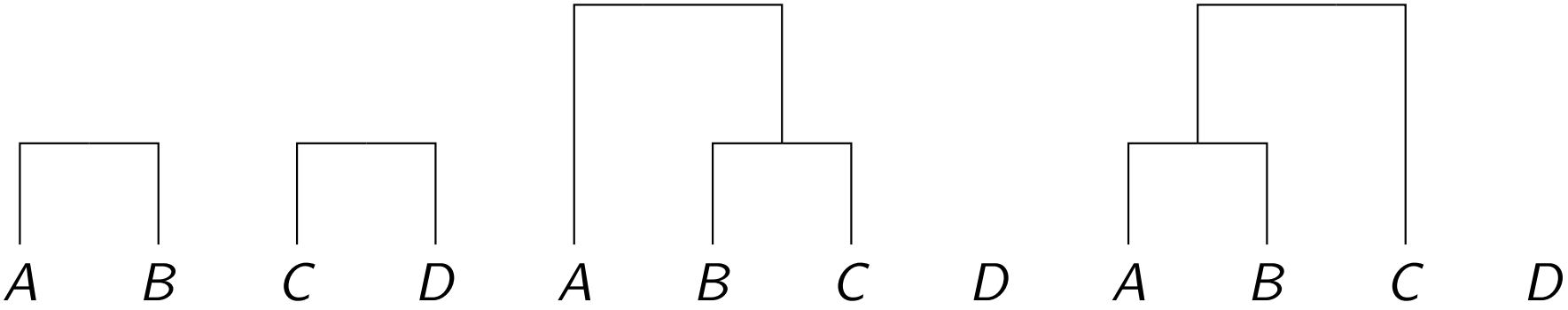




Sequential Search: Combinatorial SMC

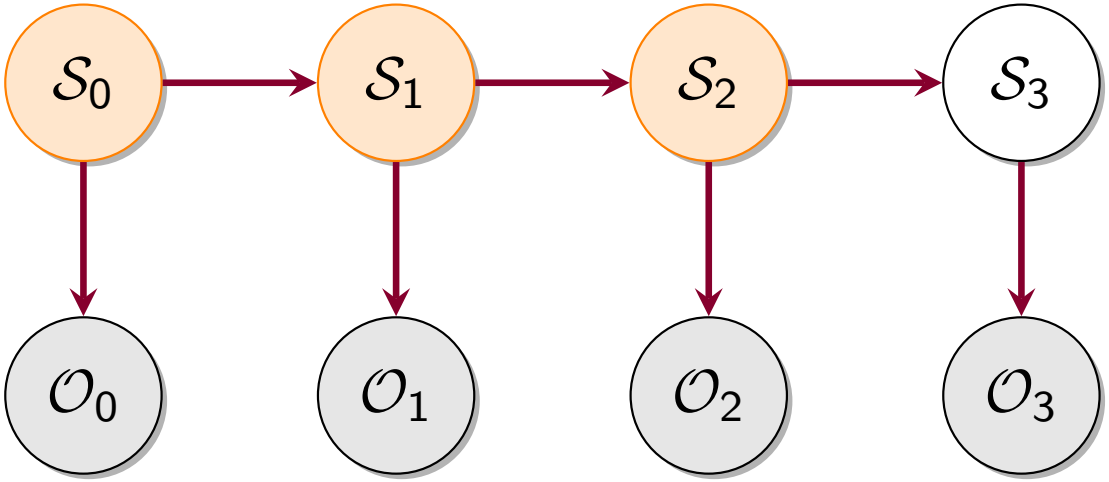
- Sample K **partial states** $\{s_{r,k}\}_{k=1}^{K} \in \mathcal{S}_r$ at each rank r

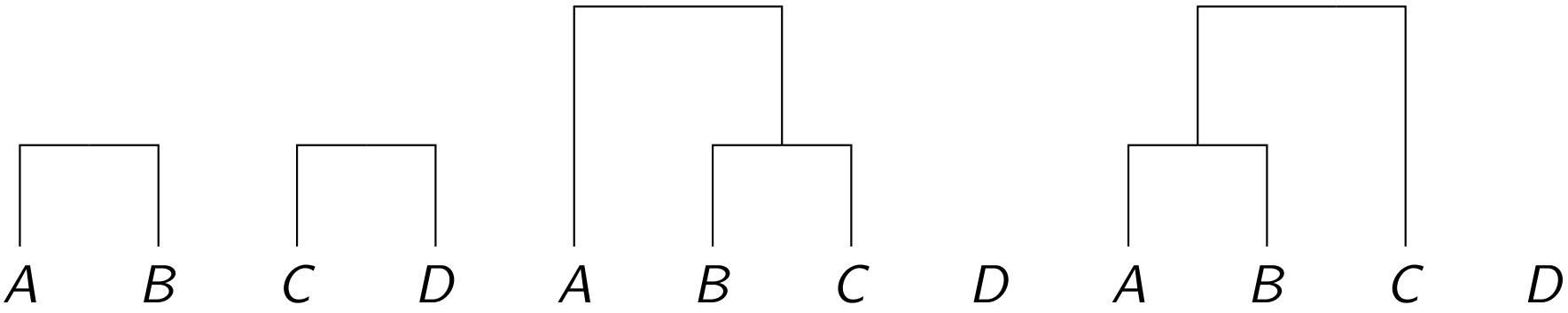




Sequential Search: Combinatorial SMC

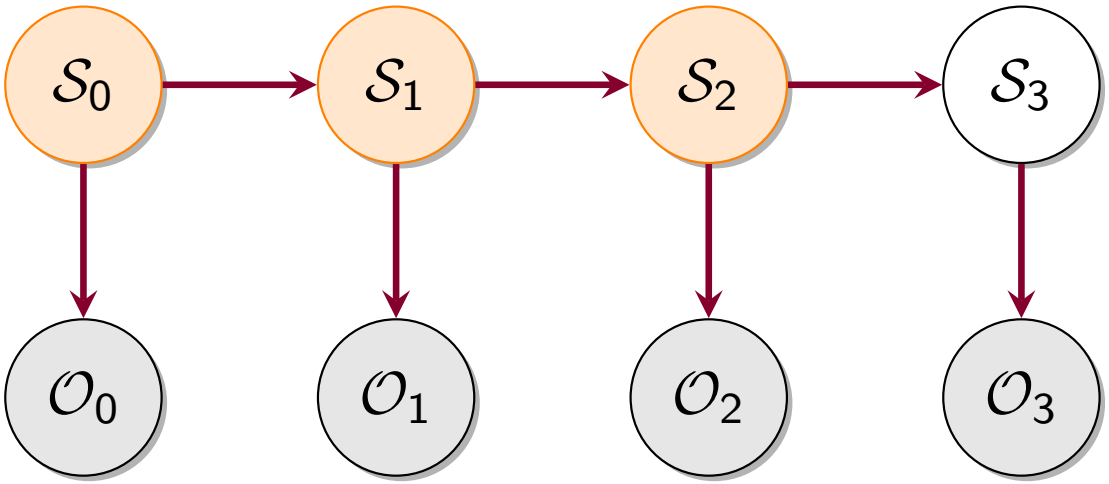
- Assign **importance weight** $\{w_{r,k}\}_{k=1}^{K}$ to each **partial state** $\{s_{r,k}\}_{k=1}^{K} \in \mathcal{S}_r$

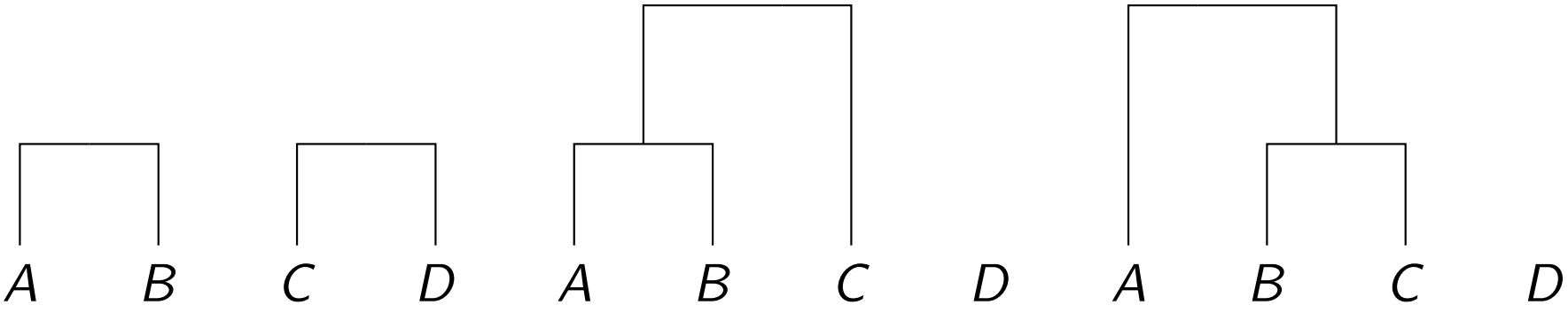




Sequential Search: Combinatorial SMC

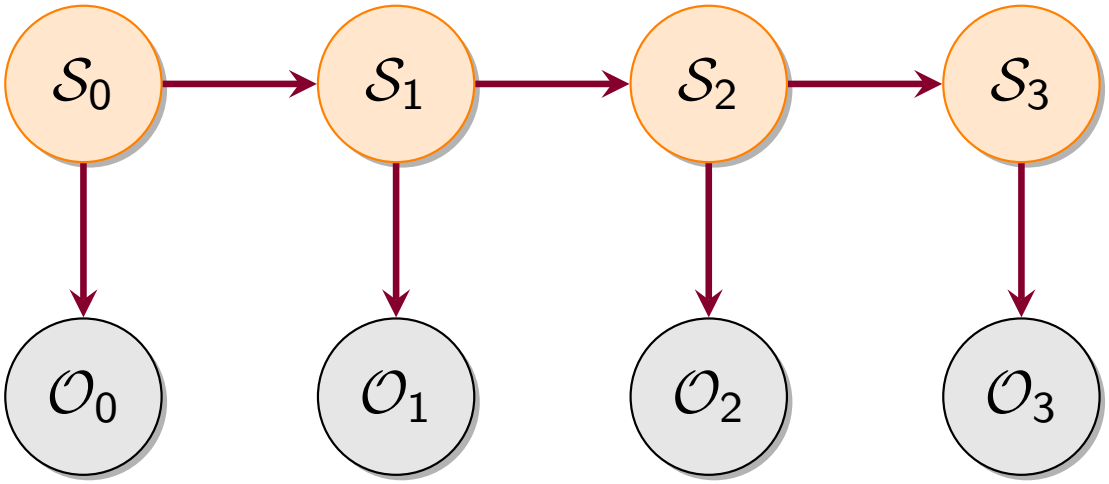
- **Resample** state $\tilde{s}_{r,k} \sim \textsc{Categorical}(\bar{w}_{r-1,1}, \cdots, \bar{w}_{r-1,K})$ to focus on areas of **high probability**.

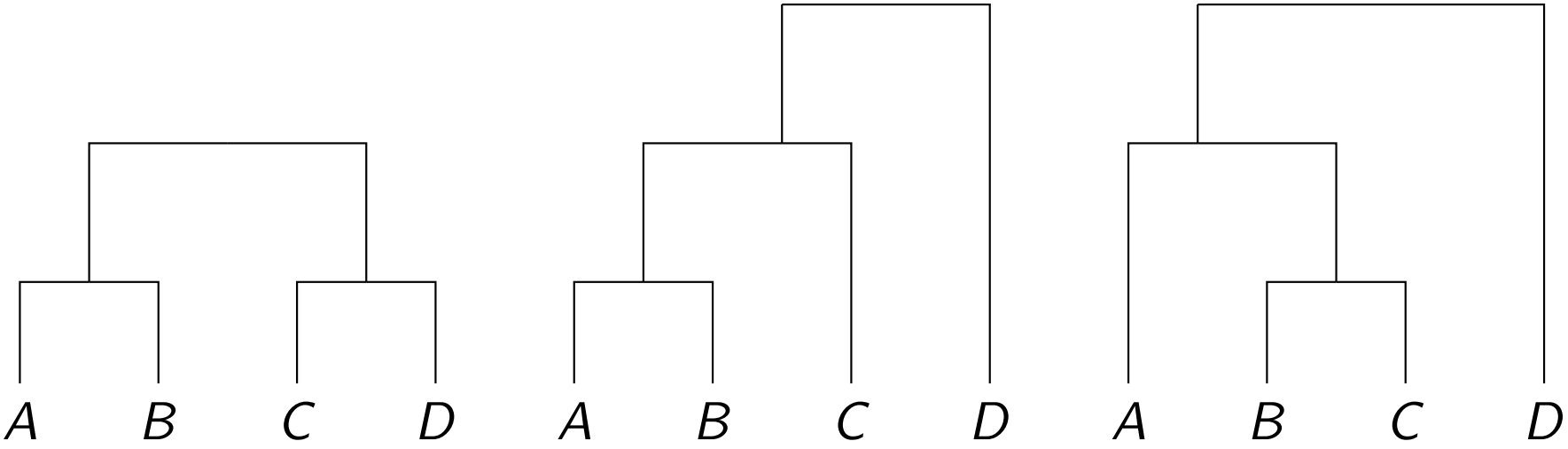




Sequential Search: Combinatorial SMC

- Sample K **partial states** $\{s_{r,k}\}_{k=1}^{K} \in \mathcal{S}_r$ at each rank r





Sequential Search: Combinatorial SMC

1. Draw K **partial states** $\{s_{r,k}\}_{k=1}^{K} \in \mathcal{S}_r$ from proposal $\nu^{+}_{s_{r,k}} : \mathcal{S} \to [0,1]$ at each rank r

$$\pi_{r,k} = \|\pi_{r-1,k}\| \frac{1}{K} \sum_{k=1}^{K} w_{r,k} \delta_{s,k}(s) \qquad \forall s \in \mathcal{S}$$

2. Compute **importance weights**

$$w_{r,k} = w_{(}\tilde{s}_{r-1,k}, s_{r,k}) = \frac{\pi(s_{r,k})}{\pi(\tilde{s}_{r-1,k})} \cdot \frac{\nu^{-}_{s_{r,k}}(\tilde{s}_{r-1,k})}{\nu^{+}_{\tilde{s}_{r,k}}(s_{r,k})},$$

3. **Resample** state $\tilde{s}_{r,k} \sim \textsc{Categorical}(\bar{w}_{r-1,1}, \cdots, \bar{w}_{r-1,K})$

$\Longrightarrow$ **Unbiased estimate** for the marginal likelihood

$$\hat{\mathcal{Z}}_{\textsc{Csmc}} := \|\pi_{R,K}\| = \prod_{r=1}^{R} \left(\frac{1}{K} \sum_{k=1}^{K} w_{r,k} \right) \to \|\pi\|.$$

Partial States and Partially Ordered Sets

Probability measure π **defined** on **target space** of *phylogenetic trees* $\mathcal{X}$, **not larger space** of *partial states* $\mathcal{S}_r$

1. Sets of partial states of different ranks disjoint:

$$\mathcal{S}_r \cap \mathcal{S}_q = \emptyset \quad \forall r \neq q$$

2. Sets of partial states of smallest rank has singleton:

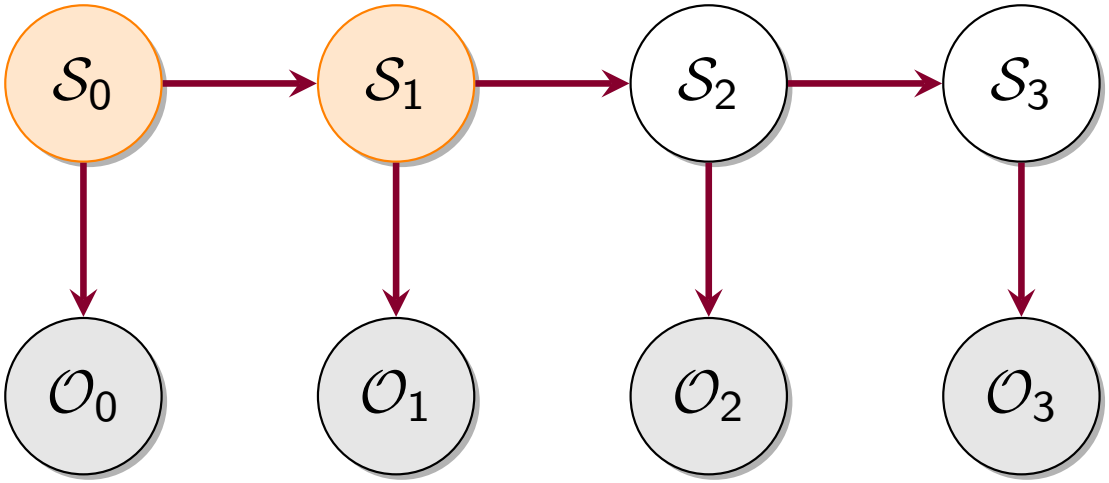
$$\mathcal{S}_0 = \{\perp\}$$

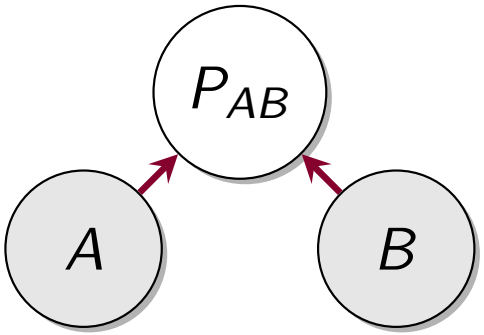
3. Set of partial state of rank R is target measure:

$$\mathcal{S}_R = \mathcal{X}$$

Extending the Target Measure

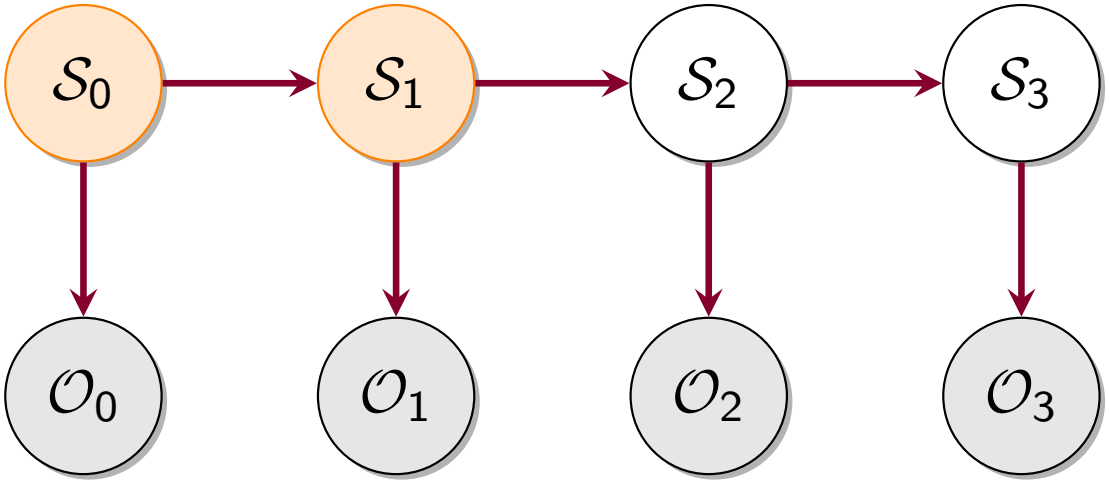
- Probability measure π **defined** on **target space** of *phylogenetic trees* $\mathcal{X}$, **not larger space** of *partial states* $\mathcal{S}_r$



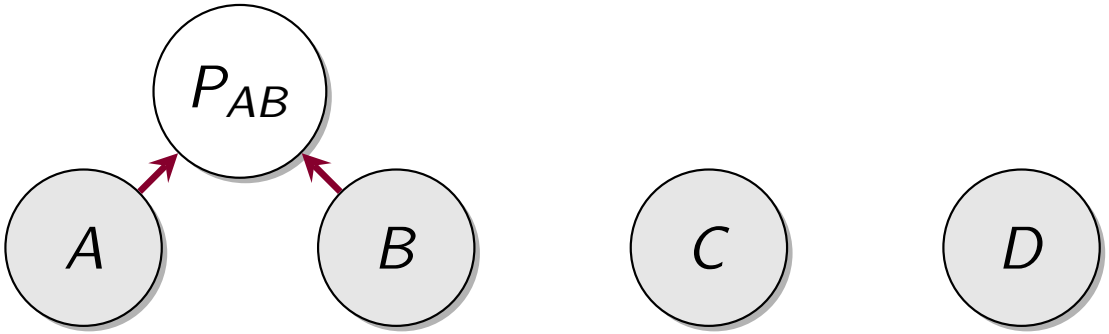


Extending the Target Measure

- Probability measure π **defined** on **target space** of *phylogenetic trees* $\mathcal{X}$, **not larger space** of *partial states* $\mathcal{S}_r$

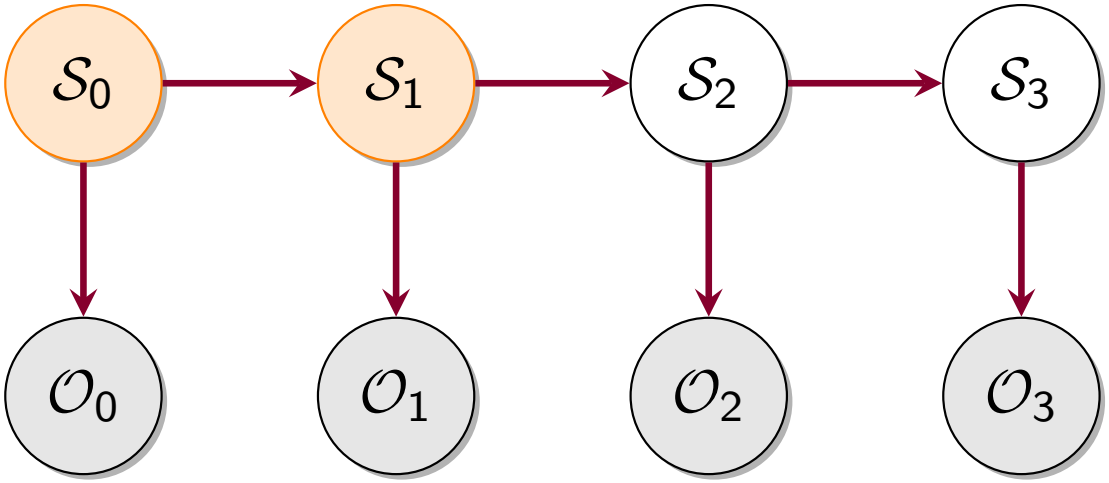


$$\pi(s) = \prod_{(t_i, X_i) \in s} \pi_{Y_i(x_i)}(t_i)$$

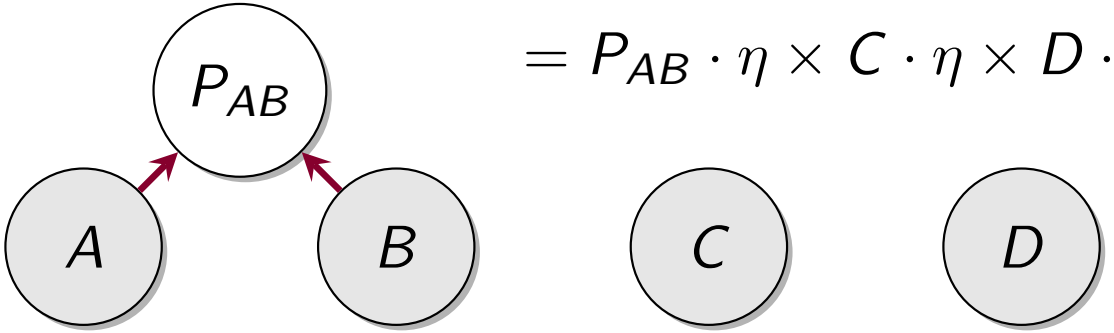


Extending the Target Measure

- Probability measure π **defined** on **target space** of *phylogenetic trees* $\mathcal{X}$, **not larger space** of *partial states* $\mathcal{S}_r$



$$\pi(s) = \prod_{(t_i, X_i) \in s} \pi_{Y_i(x_i)}(t_i)$$
$$= P_{AB} \cdot \eta \times C \cdot \eta \times D \cdot \eta$$



Variational Combinatorial Sequential Monte Carlo

Can we design **variational objective** on **composite space** of *non-clock phylogenetic trees* using sequential search?

Variational Combinatorial Sequential Monte Carlo

Can we design **variational objective** on **composite space** of *non-clock phylogenetic trees* using sequential search?

- Develop fast alternatives to MCMC for both inference and learning in Bayesian phylogenetics

Variational Combinatorial Sequential Monte Carlo

Can we design **variational objective** on **composite space** of *non-clock phylogenetic trees* using sequential search?

- Stochastic gradient VI with **variance reduction** and **reparameterization** on *discrete structures*

Variational Combinatorial Sequential Monte Carlo

Can we design **variational objective** on **composite space** of *non-clock phylogenetic trees* using sequential search?

- Use proposal $Q_\phi(\mathcal{B}, \tau | \mathsf{Y})$ to form lower bound to marginal log-evidence:

$$\log P_\theta(\mathsf{Y}) \geq \mathcal{L}_{\mathsf{ELBO}}(\theta, \phi, \mathsf{Y}) := \mathop{\mathbb{E}}_{Q}\left[\log \frac{P_\theta(\mathsf{Y}, \mathcal{B}, \tau)}{Q_\phi(\mathcal{B}, \tau | \mathsf{Y})}\right].$$

Variational Combinatorial Sequential Monte Carlo

Can we design **variational objective** on **composite space** of *non-clock phylogenetic trees* using sequential search?

- Use proposal $Q_\phi(\mathcal{B}, \tau | \mathrm{Y})$ to form lower bound to marginal log-evidence:

$$\log P_\theta(\mathrm{Y}) \geq \mathcal{L}_{\mathrm{ELBO}}(\theta, \phi, \mathrm{Y}) := \mathop{\mathbb{E}}_{Q}\left[\log \frac{P_\theta(\mathrm{Y}, \mathcal{B}, \tau)}{Q_\phi(\mathcal{B}, \tau | \mathrm{Y})}\right].$$

- Use sequential search to form objective from estimator:

$$\mathcal{L}_{VCSMC} := \mathop{\mathbb{E}}_{Q}\left[\log \hat{\mathcal{Z}}_{VCSMC}\right], \quad \hat{\mathcal{Z}}_{VCSMC} := \prod_{r=1}^{R}\left(\frac{1}{K}\sum_{k=1}^{K} w_{r,k}\right)$$

Variational Combinatorial Sequential Monte Carlo

Writing discrete ϕ and continuous ψ proposal terms explicitly:

$$
\begin{aligned}
&Q_{\phi,\psi}\left(\mathcal{S}_{1:R}^{1:K}\right) := \\
&\left(\prod_{k=1}^{K} q_\phi(s_{1,k}) \cdot q_\psi(\mathcal{B}_{1,k})\right) \times \\
&\left(\prod_{k=1}^{K}\prod_{r=1}^{N-1} q_\phi\left(s_{r,k} | s_{r-1}^{a_{r-1}^k}\right) \cdot q_\psi\left(\mathcal{B}_{r,k} | \mathcal{B}_{r-1}^{a_{r-1}^k}\right) \cdot \textsc{Cat}\left(a_{r-1}^k | \bar{w}_{r-1}^{1:K}\right)\right)
\end{aligned}
$$

Variational Combinatorial Sequential Monte Carlo

$\Longrightarrow$ Extend partial state $s_{r,k} \sim q_\phi(s_{r,k} | \tilde{s}_{r-1,k})$ by drawing two partial states to coalesce.

- Perturb uniform log-prob for each index by adding indep Gumbel dist noise, return largest two elements.
- $U \sim \textsc{Uniform}(0, 1)$, form $G = \gamma - \log(-\log U)$.
- G reparameterized as $G' = G + \gamma$.

Variational Combinatorial Sequential Monte Carlo

Do tighter variational bounds affect learning inference network?

Variational Combinatorial Sequential Monte Carlo

Do tighter variational bounds affect learning inference network?

- Reparameterization gradients of IWAE inference network decrease at rate $\mathcal{O}(1/\sqrt{K})$

Variational Combinatorial Sequential Monte Carlo

Do tighter variational bounds affect learning inference network?

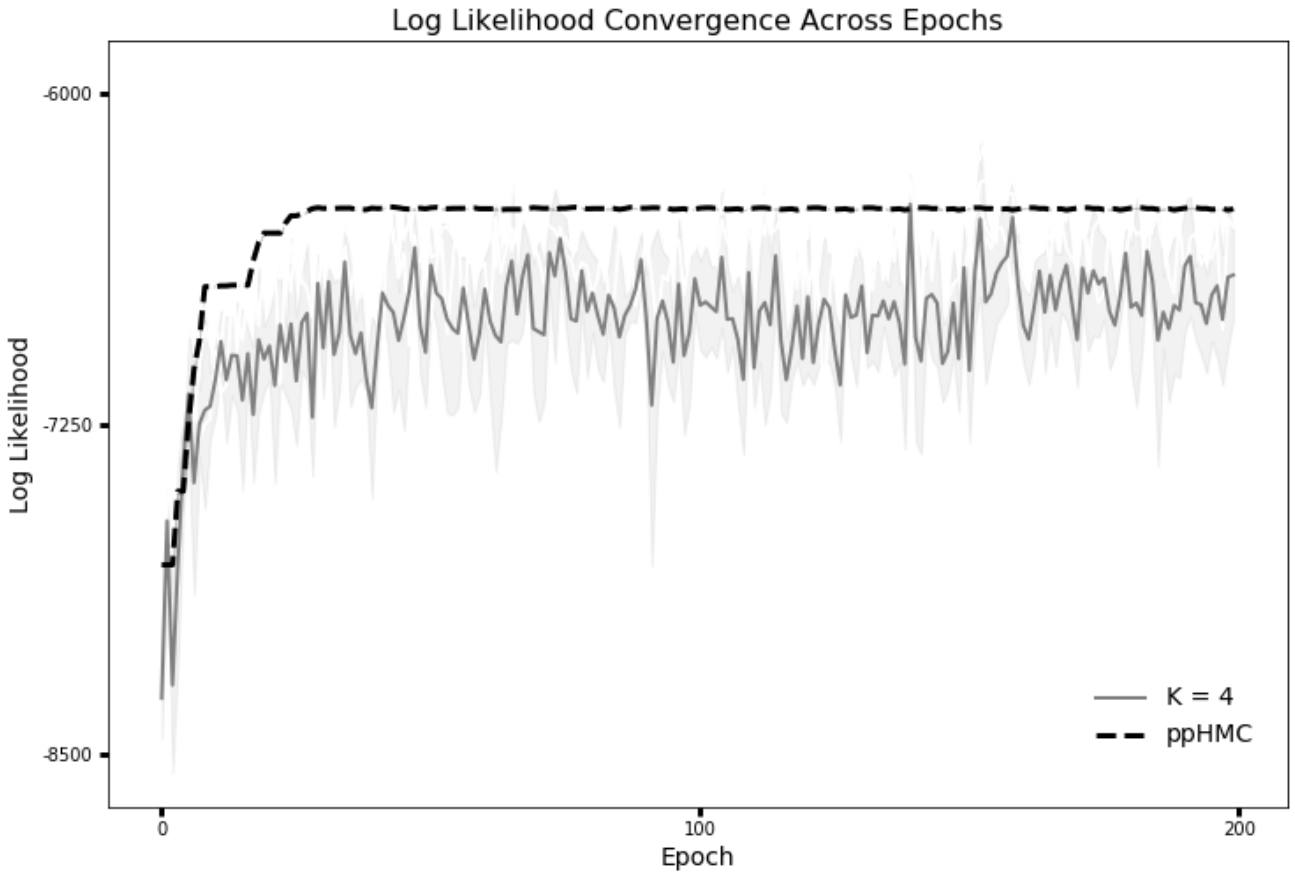
- Reparameterization gradients of IWAE inference network decrease at rate $\mathcal{O}(1/\sqrt{K})$
- VCSMC has no terms unique to inference network Q

Primate Mitochondrial DNA

Benchmark dataset:

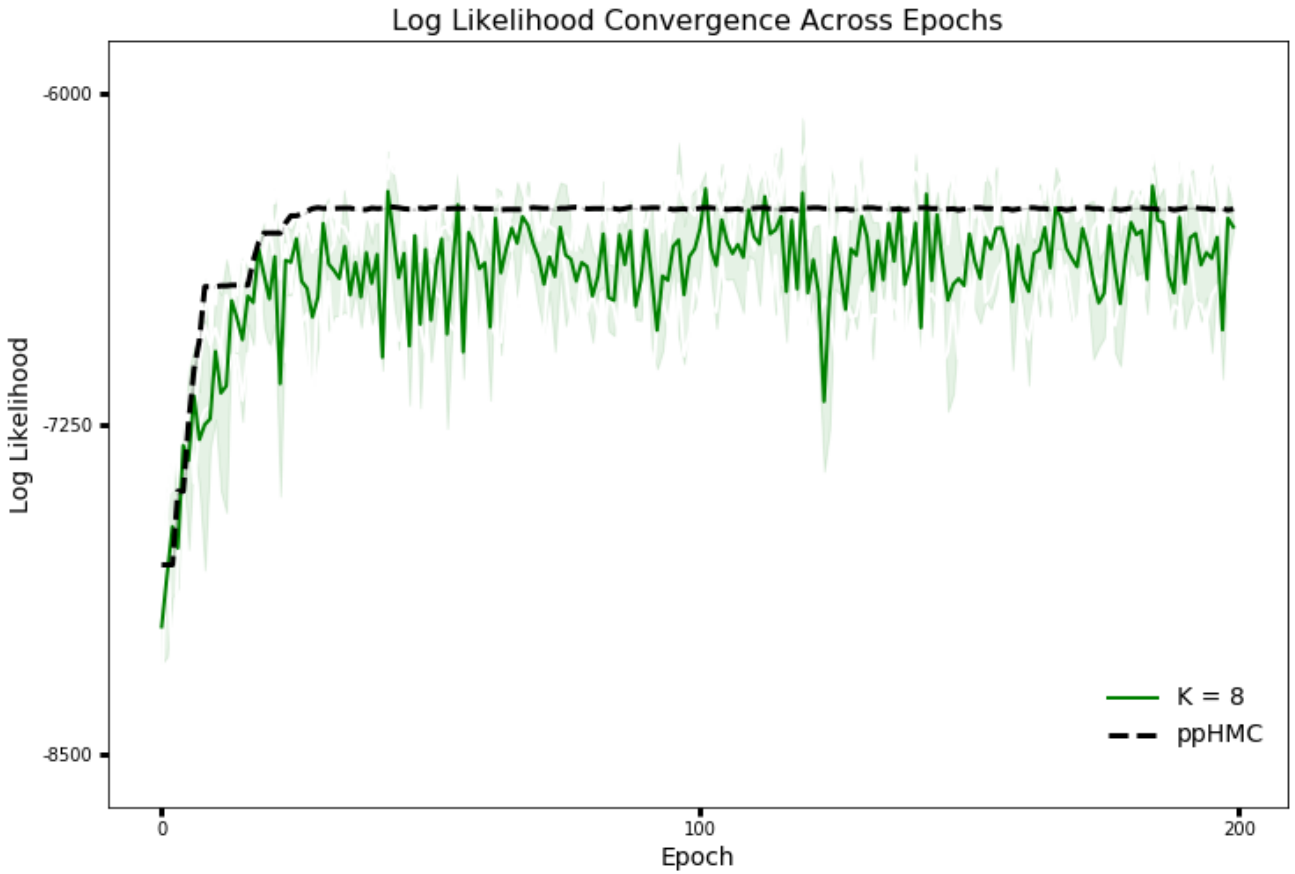
- Homologous fragments of nucleotide sequences of primate mitochondrial DNA
- 12 taxa $\{S_0, \cdots, S_{11}\}$ over 898 sites admitting 13 billion distinct topologies.
- Five homonoids, four old world monkeys, one new world monkey and two prosimians.

Primate Mitochondrial DNA



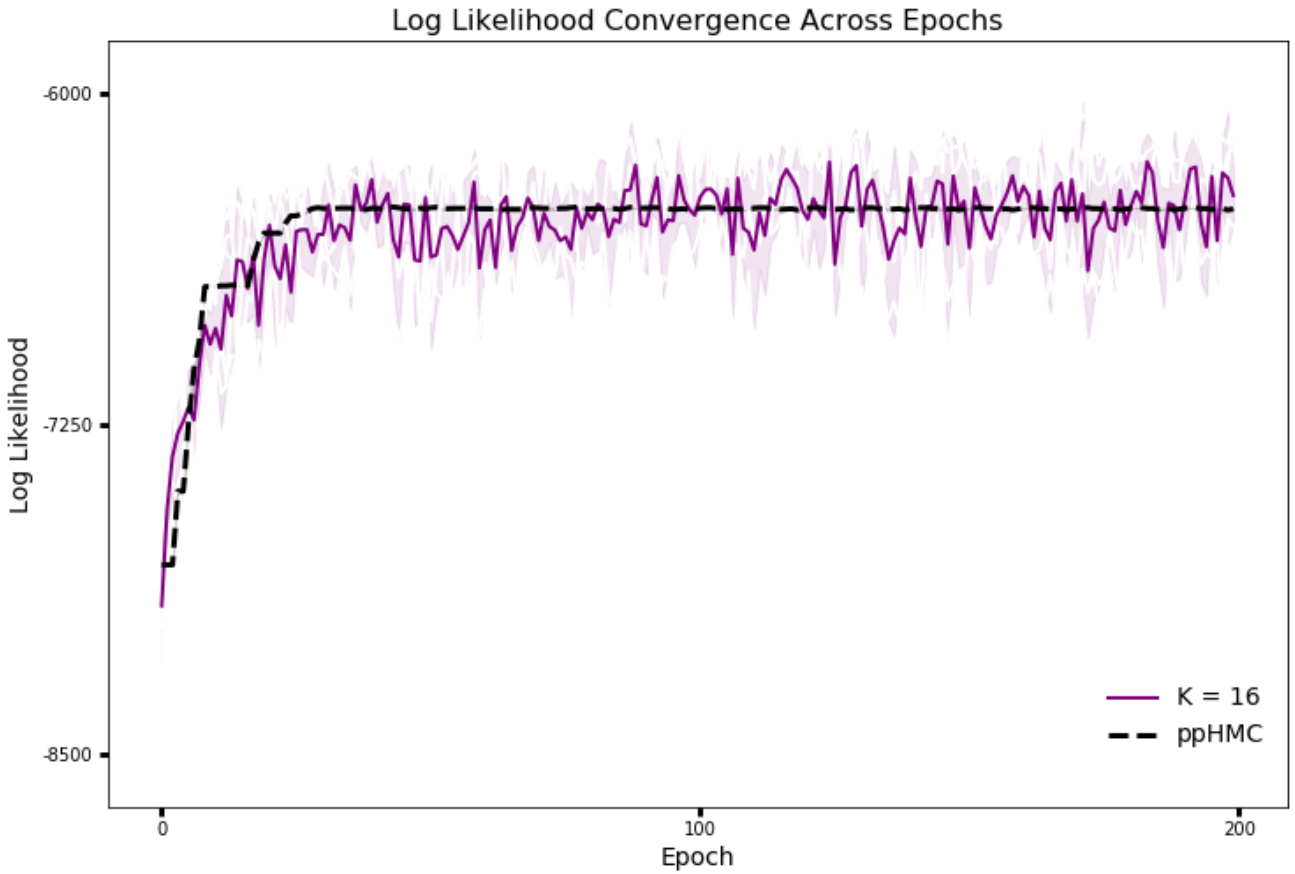
- ppHMC vs VCSMC run with $K = \{\mathbf{4}, 8, 16, 32, 64, 128\}$ samples, averaged over 3 random seeds

Primate Mitochondrial DNA



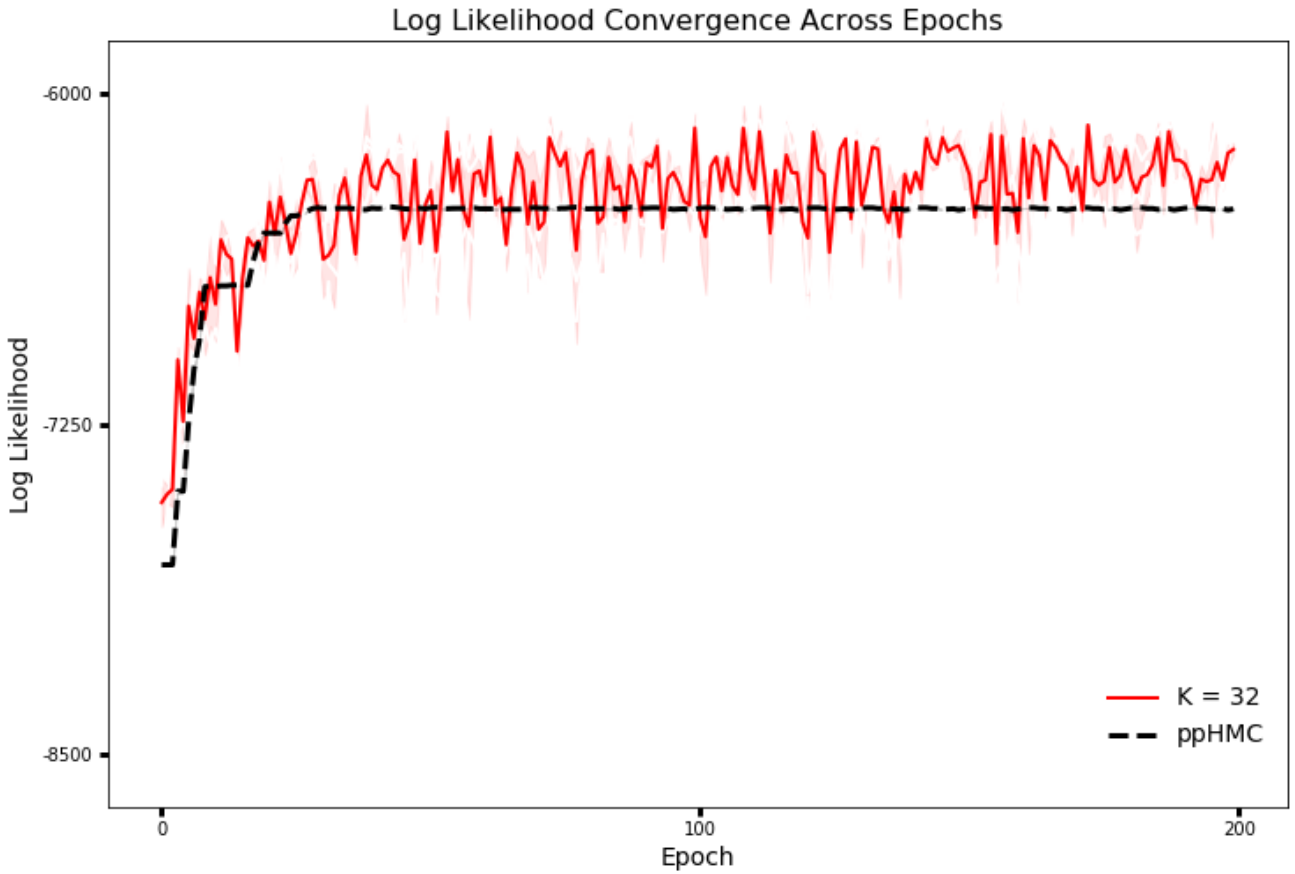
- ppHMC vs VCSMC run with $K = \{4, \mathbf{8}, 16, 32, 64, 128\}$ samples, averaged over 3 random seeds

Primate Mitochondrial DNA



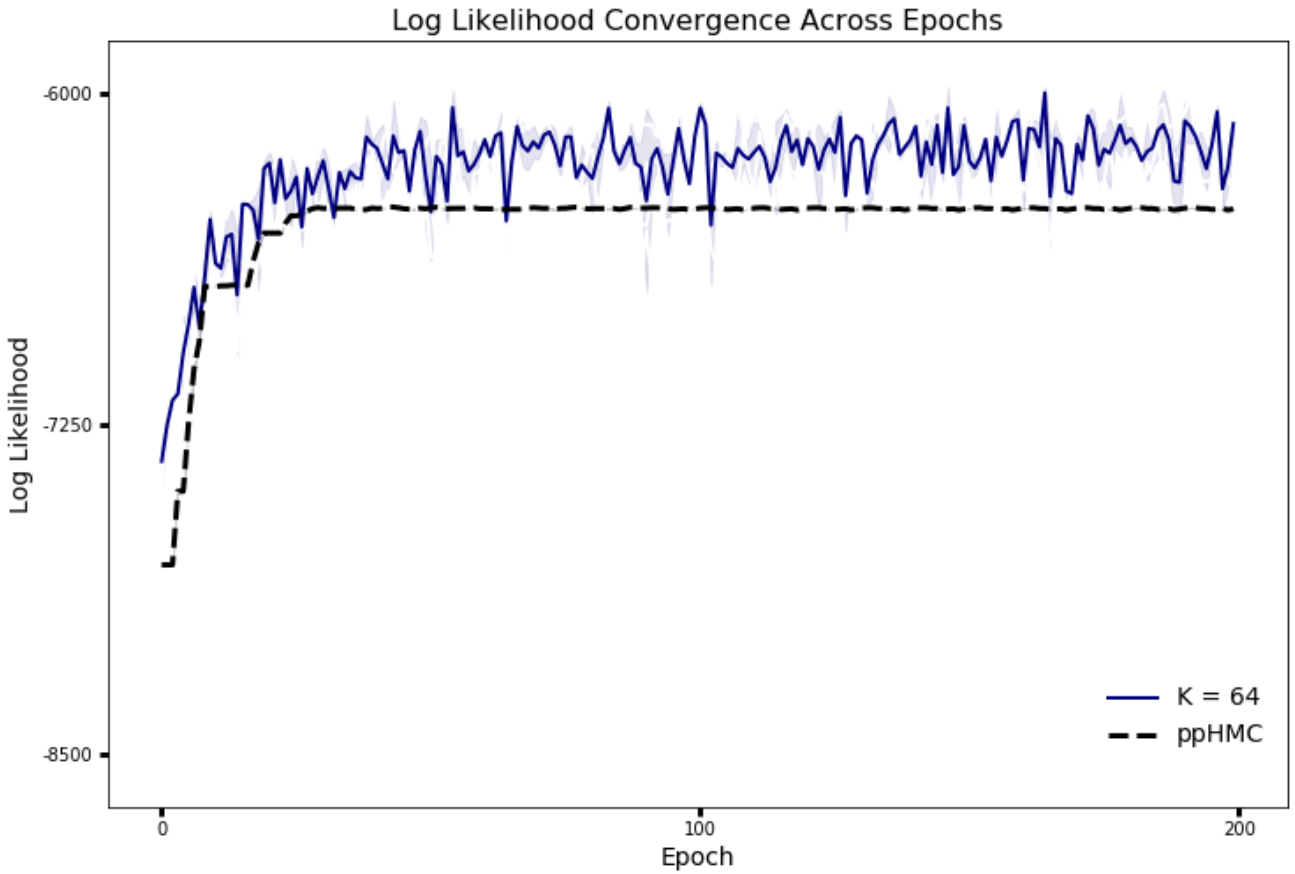
- ppHMC vs VCSMC run with $K = \{4, 8, \mathbf{16}, 32, 64, 128\}$ samples, averaged over 3 random seeds

Primate Mitochondrial DNA



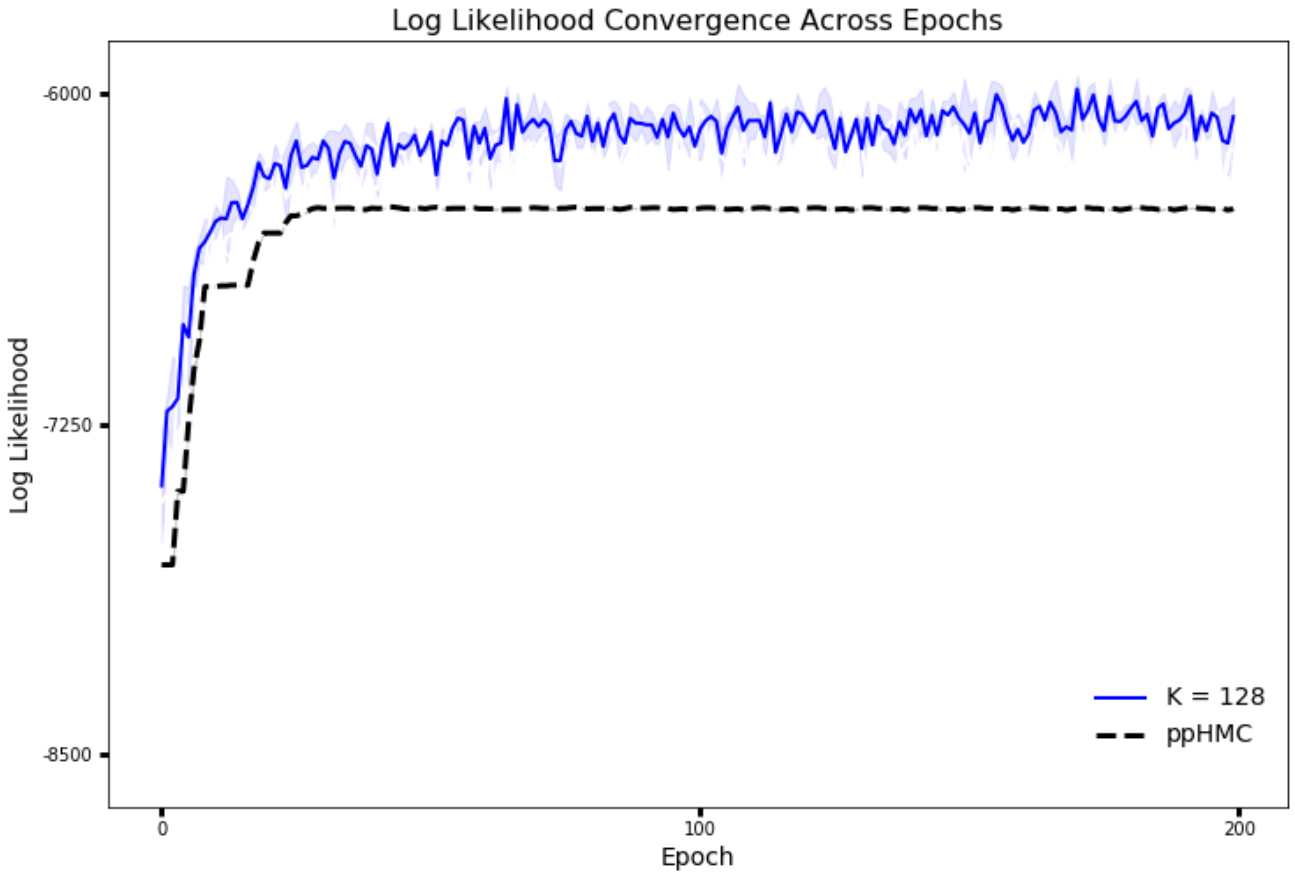
- ppHMC vs VCSMC run with $K = \{4, 8, 16, \mathbf{32}, 64, 128\}$ samples, averaged over 3 random seeds

Primate Mitochondrial DNA



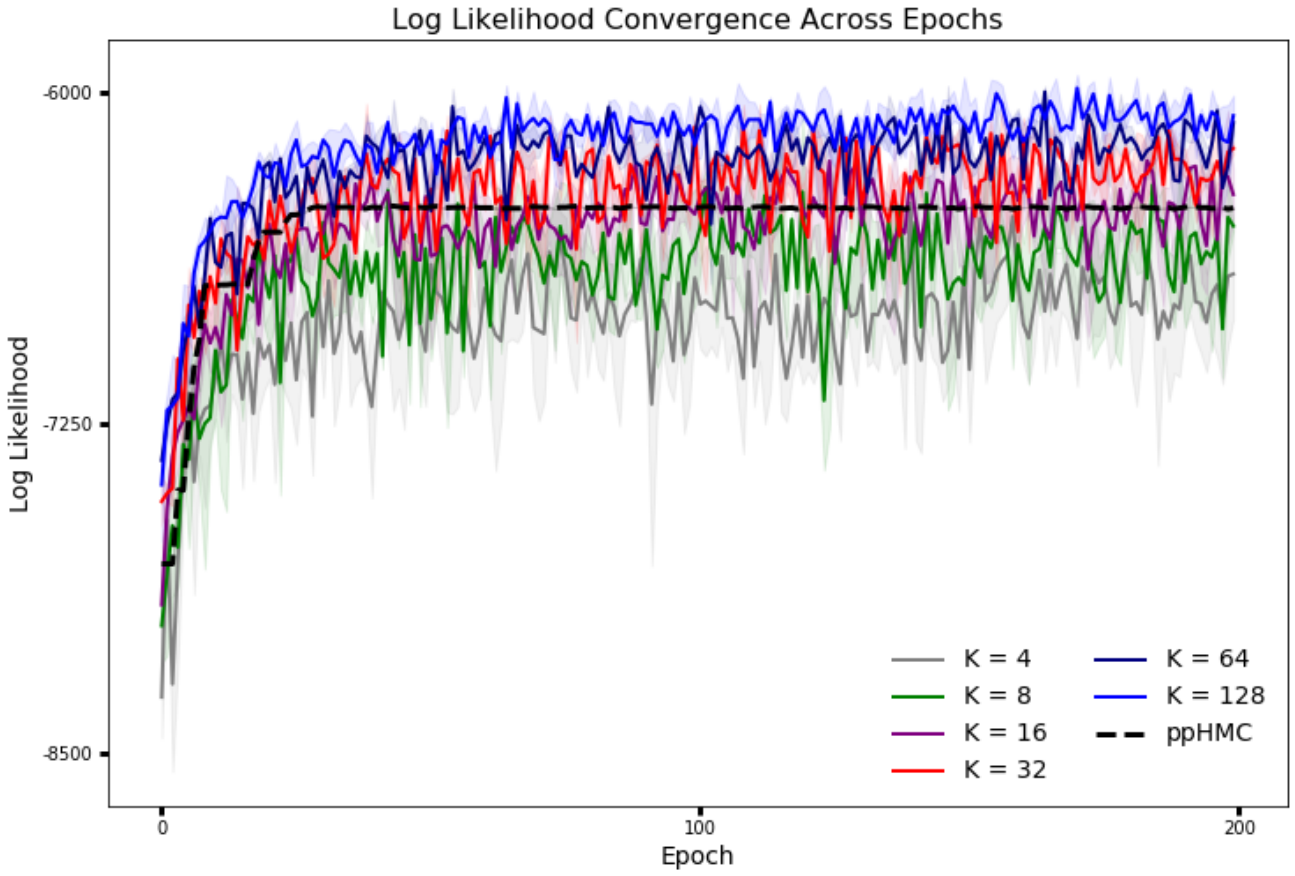
- ppHMC vs VCSMC run with $K = \{4, 8, 16, 32, \mathbf{64}, 128\}$ samples, averaged over 3 random seeds

Primate Mitochondrial DNA



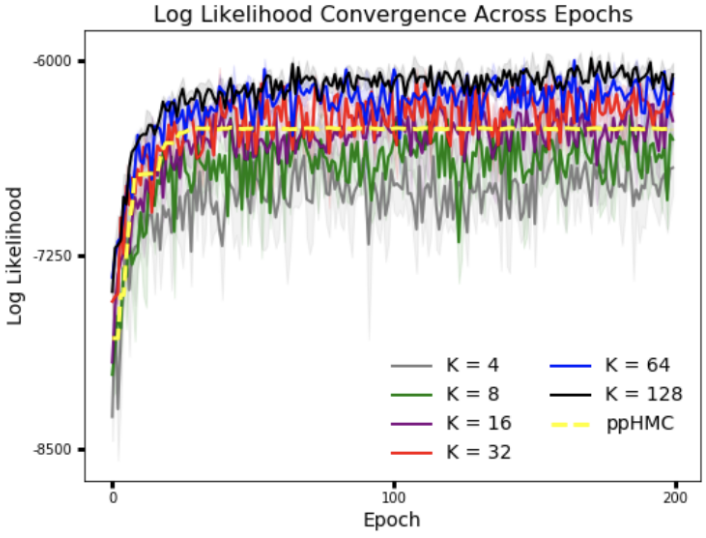
- ppHMC vs VCSMC run with $K = \{4, 8, 16, 32, 64, \mathbf{128}\}$ samples, averaged over 3 random seeds

Primate Mitochondrial DNA

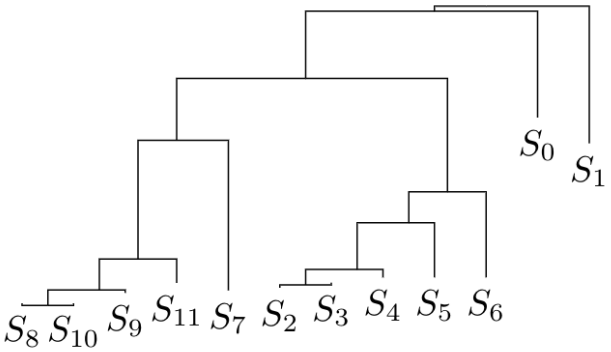


- Tighter variational bounds w/ lower stochastic gradient noise as K increases.

Primate Mitochondrial DNA



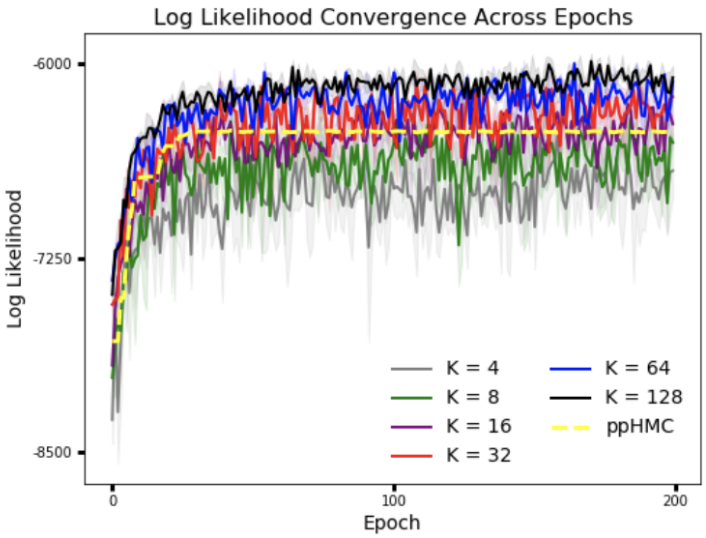
(a) Log likelihood across epochs



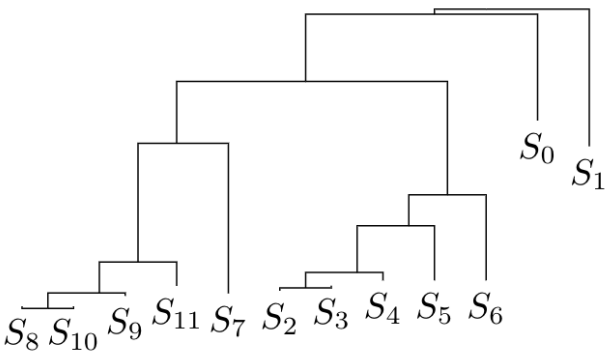
(b) Phylogeny sampled from the posterior

- Phylogeny sampled from the posterior:
  M Mulatta, M Sylvanus, M Fascicularis, Saimiri Sciureus, Macaca Fuscata, Homo Sapiens, Pan, Gorilla, Pongo, Hylobates, Tarsius Syrichta, Lemur Catta

Primate Mitochondrial DNA



(a) Log likelihood across epochs



(b) Phylogeny sampled from the posterior

- Left clade partitions **monkeys**, central and right partition **hominids** and **prosimians**.

Takeaways

VCSMC:

- VI on **composite space** of *nonclock phylogenetic trees*.

Takeaways

VCSMC:

- VI on **composite space** of *nonclock phylogenetic trees*.
- Introduces **discrete variational sequential search** to *learn distributions* over intricate combinatorial structures.

Takeaways

VCSMC:

- VI on **composite space** of *nonclock phylogenetic trees*.
- Introduces **discrete variational sequential search** to *learn distributions* over intricate combinatorial structures.
- Explores *high probability spaces* on benchmark dataset.

Questions

Thank you!

- Special thanks to Christian Naesseth for helpful discussions.
- Implementation available online:

  `https://github.com/amoretti86/phylo`

References I

Alexandre Bouchard-Côté, Sriram Sankararaman, and Michael Jordan.
Phylogenetic inference via sequential monte carlo.
Systematic biology, 61:579–93, 01 2012.

J. Rodney Brister, Danso Ako-adjei, Yiming Bao, and Olga Blinkova.
NCBI Viral Genomes Resource.
Nucleic Acids Research, 43(D1):D571–D577, 11 2014.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov.
Importance weighted autoencoders, 2015.

References II

Vu Dinh, Arman Bilge, Cheng Zhang, and Frederick A. Matsen, IV.
Probabilistic path Hamiltonian Monte Carlo.
volume 70 of *Proceedings of Machine Learning Research*, pages 1009–1018, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

J Felsenstein.
Evolutionary trees from dna sequences: a maximum likelihood approach.
Journal of Molecular Evolution, 17(6):368–376, 1981.

K Hayasaka, T Gojobori, and S Horai.
Molecular phylogeny and evolution of primate mitochondrial DNA.
Molecular Biology and Evolution, 5(6):626–644, 11 1988.

References III

Daniel Hernandez, Antonio Moretti, Ziqiang Wei, S. Saxena, John Cunningham, and Liam Paninski.
A novel variational family for hidden nonlinear markov models.
CoRR, abs/1811.02459, 2018.

Daniel Hernandez, Antonio Khalil Moretti, Ziqiang Wei, Shreya Saxena, John Cunningham, and Liam Paninski.
Nonlinear evolution via spatially-dependent linear dynamics for electrophysiology and calcium data.
Neurons, Behavior, Data analysis and Theory, 2018.

John P. Huelsenbeck and Fredrik Ronquist.
MRBAYES: Bayesian inference of phylogenetic trees .
Bioinformatics, 17(8):754–755, 08 2001.

References IV

Sebastian Höhna and Alexei Drummond.
Guided tree topology proposals for bayesian phylogenetic inference.
Systematic biology, 61:1–11, 01 2012.

Diederik P Kingma and Max Welling.
Auto-encoding variational bayes, 2013.

Clemens Lakner, Paul van der Mark, John P. Huelsenbeck, Bret Larget, and Fredrik Ronquist.
Efficiency of Markov Chain Monte Carlo Tree Proposals in Bayesian Phylogenetics.
Systematic Biology, 57(1):86–103, 02 2008.

Tuan Anh Le, Maximilian Igl, Tom Rainforth, Tom Jin, and Frank Wood.
Auto-encoding sequential monte carlo.
In *International Conference on Learning Representations*, 2018.

References V

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh.
The concrete distribution: A continuous relaxation of discrete random variables, 2016.

Antonio Moretti, Andrew Stirn, Gabriel Marks, and Itsik Pe'er.
Autoencoding topographic factors.
Journal of Computational Biology, 26(6):546–560, 2019.

Antonio K Moretti, Zizhao Wang, Luhuan Wu, and Itsik Pe'er.
Smoothing nonlinear variational objectives with sequential monte carlo.
ICLR Workshops, 2019.

Antonio Khalil Moretti, Zizhao Wang, Luhuan Wu, Iddo Drori, and Itsik Pe'er.
Particle smoothing variational objectives.
CoRR, abs/1909.09734, 2019.

References VI

Antonio Khalil Moretti, Zizhao Wang, Luhuan Wu, Iddo Drori, and Itsik Pe'er.
Variational objectives for markovian dynamics with backward simulation.
European Conference on Artificial Intelligence, 2020.

D.A. Morrison.
Multiple sequence alignment for phylogenetic purposes.
Aust. Syst. Bot., 19:476–539, 01 2006.

Christian Naesseth, Scott Linderman, Rajesh Ranganath, and David Blei.
Variational sequential monte carlo.
volume 84 of *Proceedings of Machine Learning Research*, pages 968–977, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

References VII

Fredrik Ronquist, Maxim Teslenko, Paul Mark, Daniel Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc Suchard, and John Huelsenbeck.
Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space.
Systematic biology, 61:539–42, 03 2012.

Charles Semple and Mike Steel.
Phylogenetics.
2003.

Liangliang Wang, Alexandre Bouchard-Côté, and Arnaud Doucet.
Bayesian phylogenetic inference using a combinatorial sequential monte carlo method.
Journal of the American Statistical Association, 01 2015.

References VIII

Shijia Wang and Liangliang Wang.
Particle gibbs sampling for bayesian phylogenetic inference, 2020.

Cheng Zhang and Frederick A Matsen IV.
Generalizing tree probability estimation via bayesian networks.
In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1444–1453. Curran Associates, Inc., 2018.

Cheng Zhang and Frederick A Matsen IV.
Variational bayesian phylogenetic inference.
In *International Conference on Learning Representations*, 2019.