Generalizing Tree Probability Estimation via Bayesian Networks

Cheng Zhang and Frederick A. Matsen IV

czhang23@fredhutch.org Program in Computational Biology, Fred Hutchinson Cancer Research Center

Introduction

Goal: Estimate the probability of phylogenetic (i.e. evolutionary) trees based on MCMC samples

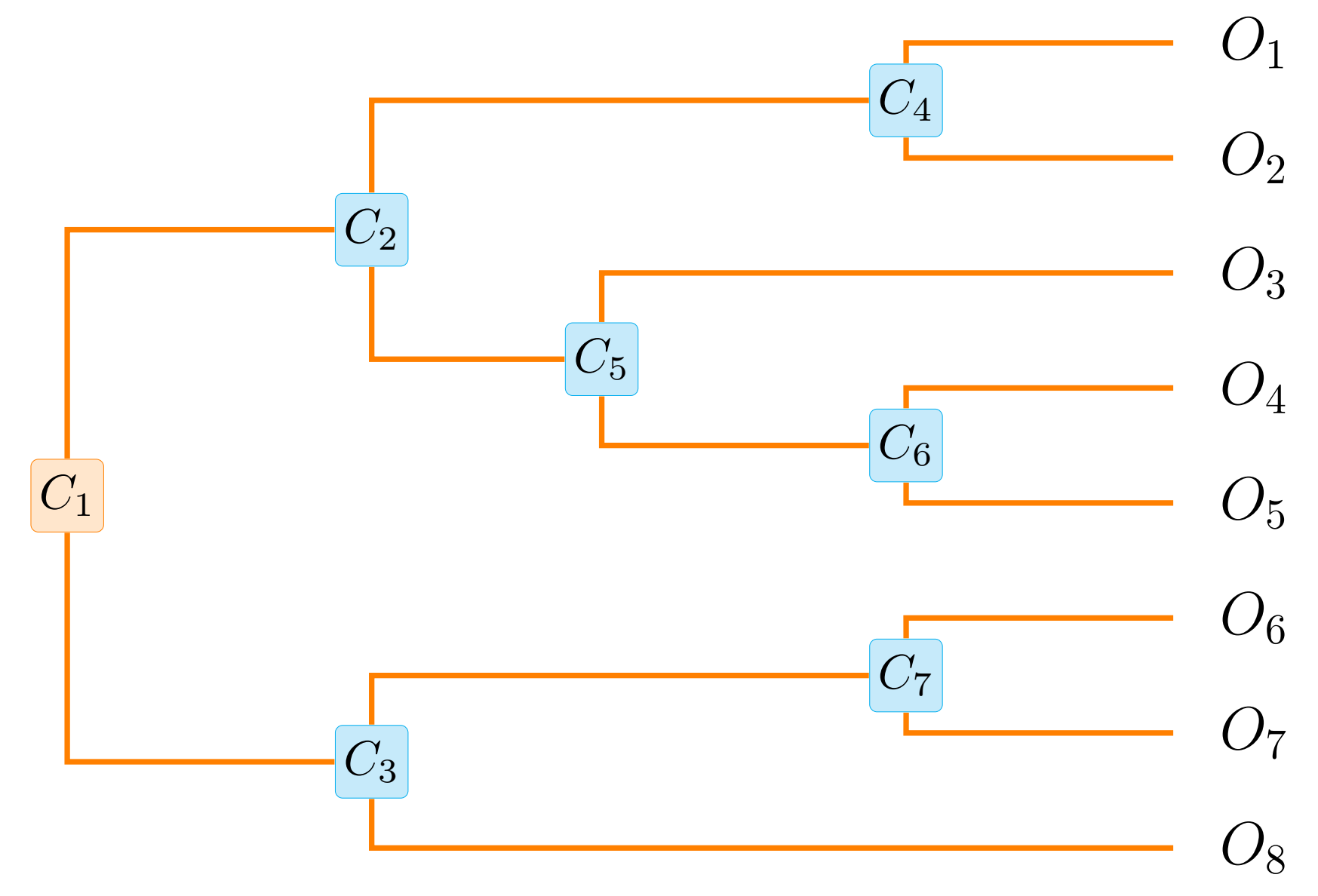
Motivation: Current methods are unsatisfactory

- The common practice of using simple sample relative frequencies (SRF) does not support unsampled trees, and is prone to large variance between different runs
- Previous efforts do extend to unsampled trees, but make too strong assumptions to provide accurate posterior estimation for real data.

By introducing a novel graphical model, **subsplit Bayesian networks**, we propose a general probability estimation framework for phylogenetic trees that

- generalizes to unsampled trees
- provides accurate approximation for real data posteriors

Problem Setup



A phylogenetic tree T is a binary tree with labeled leaves.

- label set $\mathcal{X} = \{O_1, \ldots, O_N\}$, each label represents a species.
- A *clade* X is a nonempty subset of $\mathcal{X}$

Conditional Clade Distribution

- Clade Decomposition (follow the splitting process of the tree).

$$T_{\mathcal{C}} = \{C_2, C_3, C_4, C_5, C_6, C_7\}$$

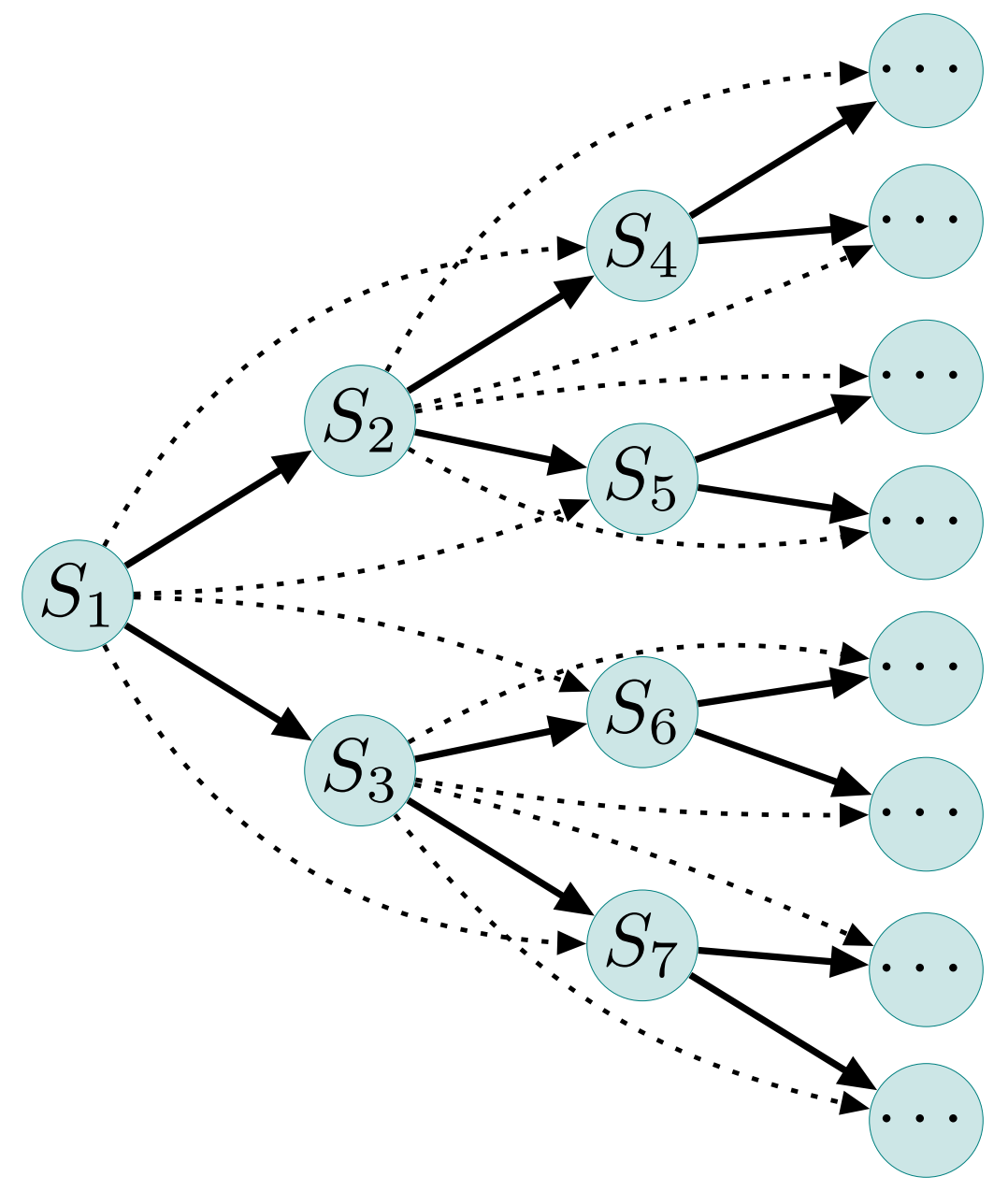
- Conditional Independent Approximation

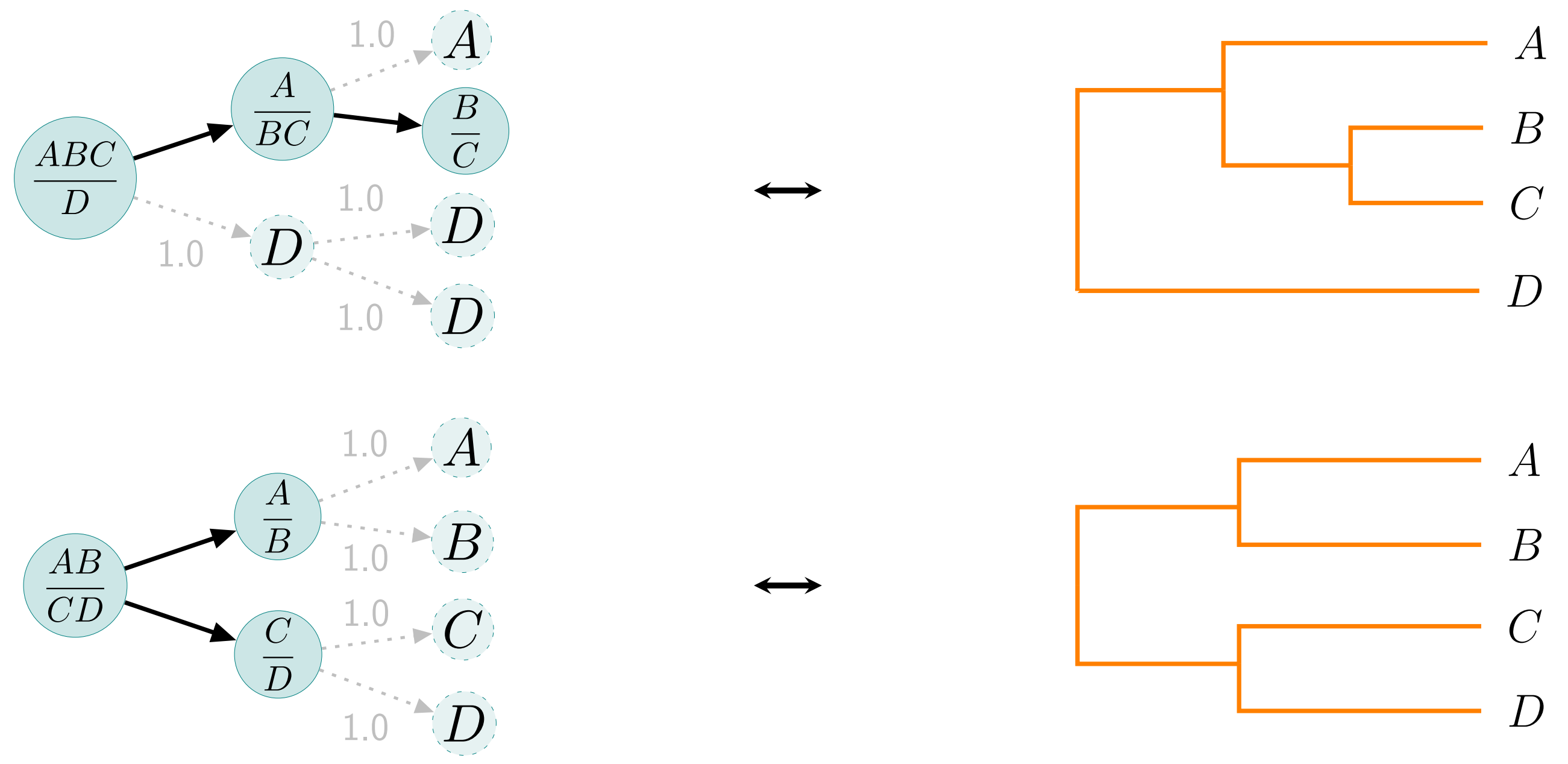
$$\begin{aligned} p_{\mathrm{ccd}}(T) &= p(C_2, C_3, C_4, C_5, C_6, C_7) \\ &= p(C_2, C_3)p(C_4, C_5|C_2)p(C_6|C_5)p(C_7|C_3) \end{aligned}$$

Let $\succ$ be a total order on clades. A *subsplit* (Y, Z) of a clade X is an ordered pair of disjoint subclades of X such that $Y \cup Z = X,\ Y \succ Z$.

- Subsplit Decomposition

$$T_{\mathcal{S}} = \{(C_2, C_3), (C_4, C_5), (\{O_3\}, C_6), (C_7, \{O_8\})\}$$





Subsplit Bayesian Network

A **subsplit Bayesian network** (SBN) $\mathcal{B}_{\mathcal{X}}$ on a leaf set $\mathcal{X}$ of size N is a Bayesian network of depth $N-1$ whose nodes take on subsplit or singleton clade values of $\mathcal{X}$ and

- the root node takes on subsplits of the entire leaf set $\mathcal{X}$
- contains a full and complete binary tree $\mathcal{B}^*_{\mathcal{X}}$

SBNs probability for rooted trees

$$p_{\mathrm{sbn}}(T) = p(S_1) \prod_{i>1} p(S_i|S_{\pi_i})$$

SBNs provide *valid probability distributions* of the entire tree space and are *flexible* to capture complicated dependence structures.

ML for Rooted Trees

Given a sample of rooted trees $\mathcal{D} = \{T_k\}_{k=1}^K$, where $T_k = \{S_i = s_{i,k},\ i \geq 1\},\ k = 1, \ldots, K$, the SBN log-likelihood function is

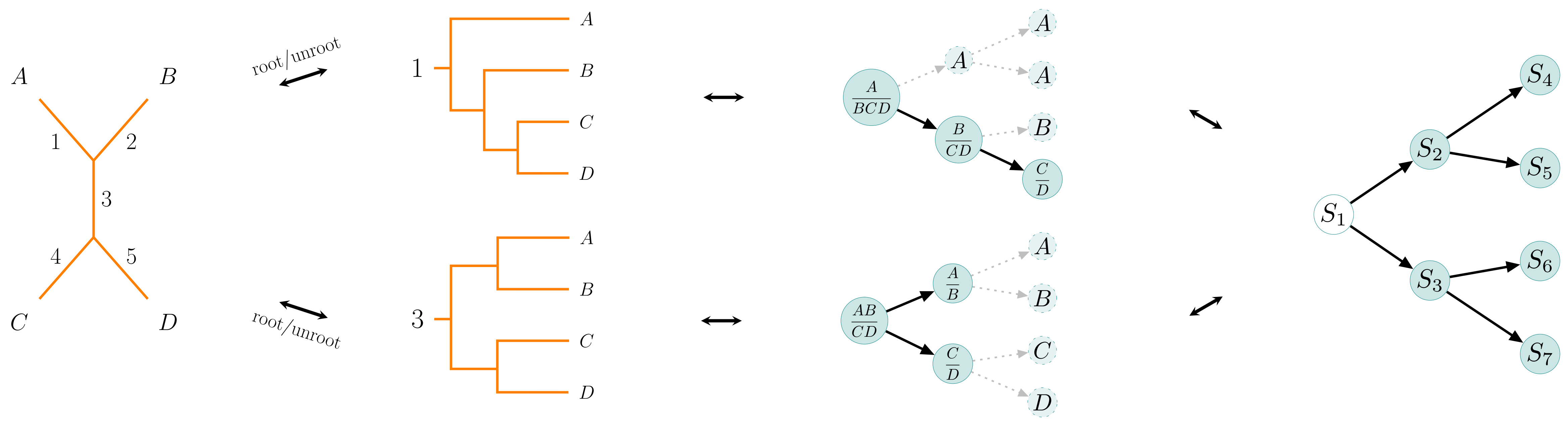
$$\log L(\mathcal{D}) = \sum_{k=1}^K \left(\log p(S_1 = s_{1,k}) + \sum_{i>1} \log p(S_i = s_{i,k} | S_{\pi_i} = s_{\pi_i,k}) \right)$$

Using *conditional probability sharing*, we have

$$\log L(\mathcal{D}) = \sum_{s_1 \in \mathbb{C}_r} m_{s_1} \log p(S_1 = s_1) + \sum_{s|t \in \mathbb{C}_{\mathrm{ch|pa}}} m_{s,t} \log p(s|t)$$

where $\mathbb{C}_r$ denotes the set of all observed root splits of S_1, $\mathbb{C}_{\mathrm{ch|pa}}$ denotes the set of all observed parent-child subsplit pairs, and $m_{s_1}, m_{s,t}$ denotes the corresponding frequency counts.

Learning SBNs for Unrooted Trees



- **Root Marginalization**

$$p_{\mathrm{sbn}}(T^{\mathrm{u}}) = \sum_{S_1 \sim T^{\mathrm{u}}} p(S_1) \prod_{i>1} p(S_i|S_{\pi_i})$$

where $\sim$ means all root subsplits that are compatible with T^{u}.

- **Variational Lower Bounds**

$$L_q(T^{\mathrm{u}}) = \sum_{S_1 \sim T^{\mathrm{u}}} q(S_1) \log \frac{p(S_1) \prod_{i>1} p(S_i|S_{\pi_i})}{q(S_1)} \leq \log p_{\mathrm{sbn}}(T^{\mathrm{u}})$$

where q is a probability distribution on $S_1 \sim T^{\mathrm{u}}$.

Lower Bounds Maximization

- **Simple Averaging**

$$q(S_1 = s_1) = \frac{1}{2N-3},\ \forall s_1 \in \mathbb{C}_r$$

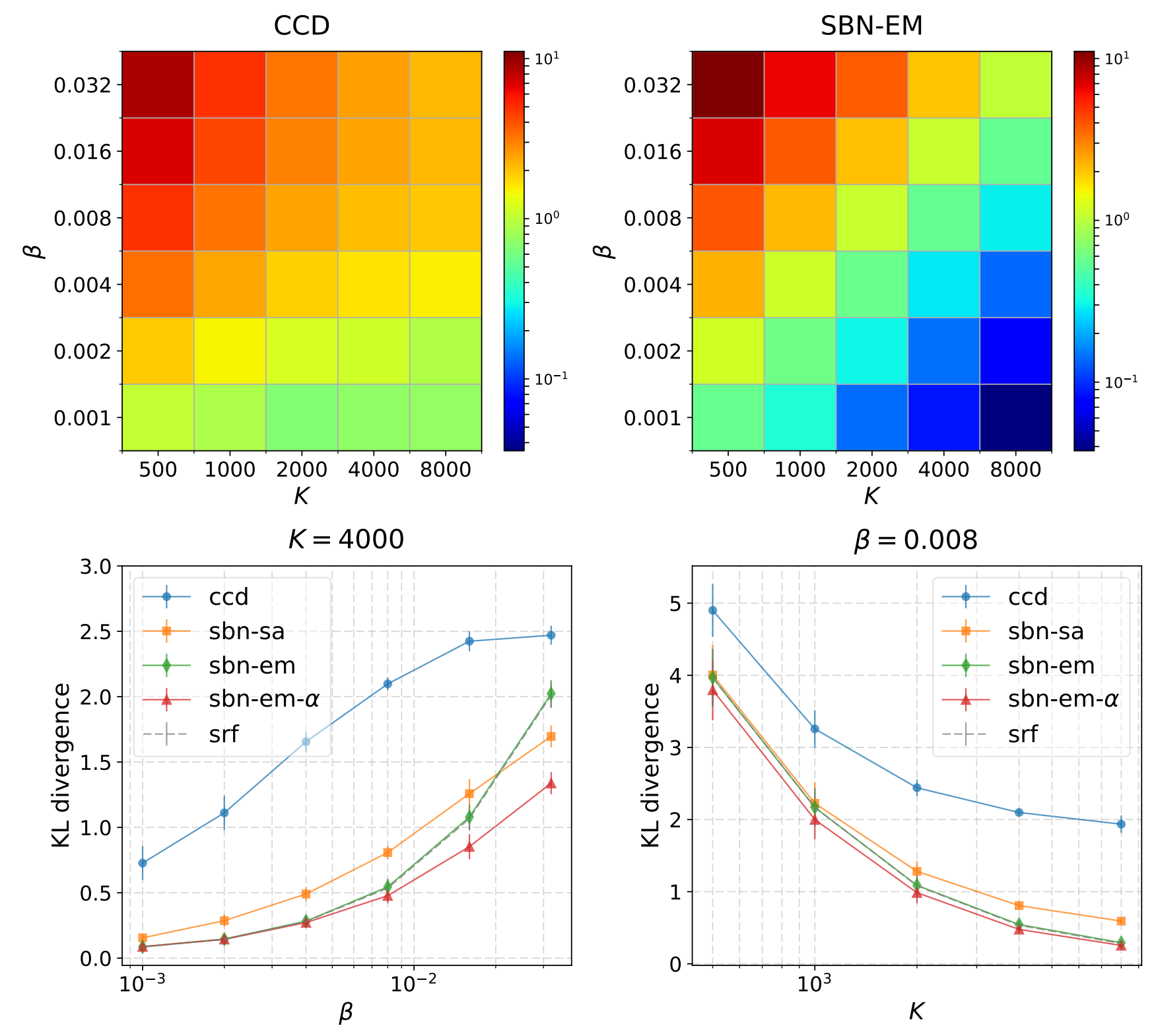
- **Expectation Maximization**

$$q^{(n)}(S_1 = s_1) = p(S_1 = s_1 | T^{\mathrm{u}}, \hat{p}^{\mathrm{EM},(n)}),\ \forall s_1 \in \mathbb{C}_r$$

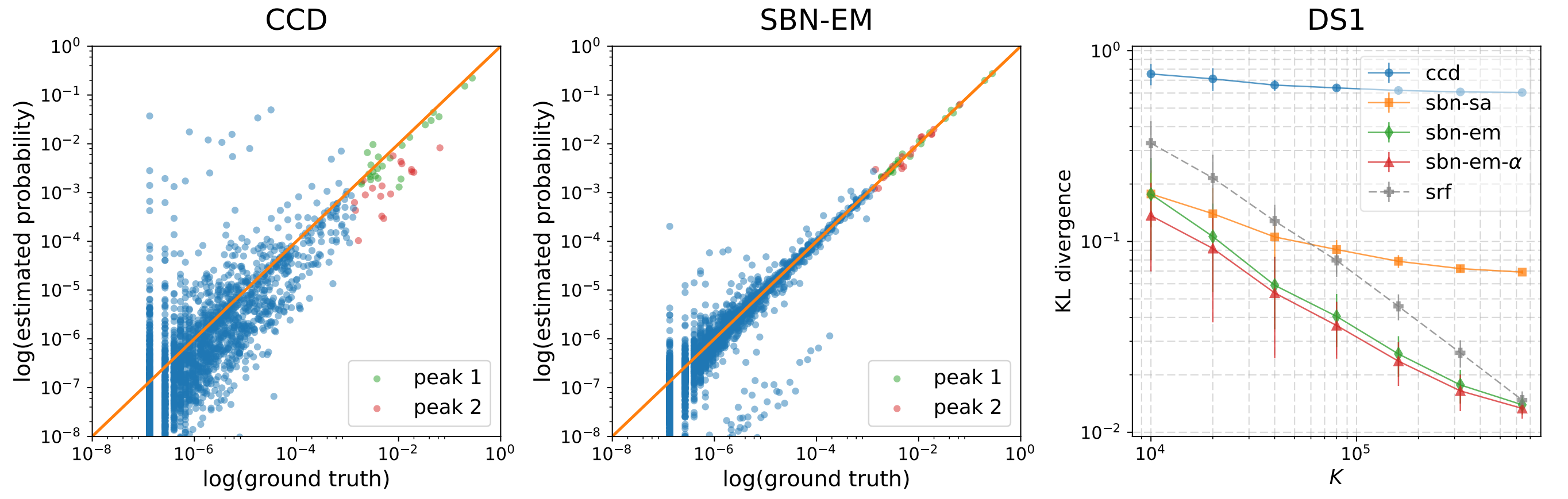
Remark: can incorporate regularization when data is insufficient or the number of parameters is large.

Experiments

SBN algorithms perform consistently much better than CCD on a challenging tree probability estimation problem with simulated data.



SBN algorithms relax the conditional clade independence assumption and provides accurate approximation in multimodal distributions.



When applied to a broad range of data sets, we find that SBNs consistently outperform other methods.

Data set	(#Taxa, #Sites)	Tree space size	Sampled trees	KL divergence to ground truth				
				SRF	CCD	SBN-SA	SBN-EM	SBN-EM-α
DS1	(27, 1949)	5.84×10^{32}	1228	0.0155	0.6027	0.0687	0.0136	**0.0130**
DS2	(29, 2520)	1.58×10^{35}	7	**0.0122**	0.0218	0.0218	0.0199	0.0128
DS3	(36, 1812)	4.89×10^{47}	43	0.3539	0.2074	0.1152	0.1243	**0.0882**
DS4	(41, 1137)	1.01×10^{57}	828	0.5322	0.1952	0.1021	0.0763	**0.0637**
DS5	(50, 378)	2.84×10^{74}	33752	11.5746	1.3272	0.8952	0.8599	**0.8218**
DS6	(50, 1133)	2.84×10^{74}	35407	10.0159	0.4526	**0.2613**	0.3016	0.2786
DS7	(59, 1824)	4.36×10^{92}	1125	1.2765	0.3292	0.2341	0.0483	**0.0399**
DS8	(64, 1008)	1.04×10^{103}	3067	2.1653	0.4149	0.2212	0.1415	**0.1236**

Conclusion

We have proposed a general framework for tree probability estimation base on subsplit Bayesian networks. SBNs allows us to exploit the similarity among trees to provide a wide range of flexible probability estimators that generalize beyond observations. Numerical results demonstrate the importance of being both flexible and generalizing when estimating probabilities on trees. We hope that these ideas will help practitioners design more efficient tree proposals for MCMC transition kernels and inspire new structural learning methods for phylogenetic models.