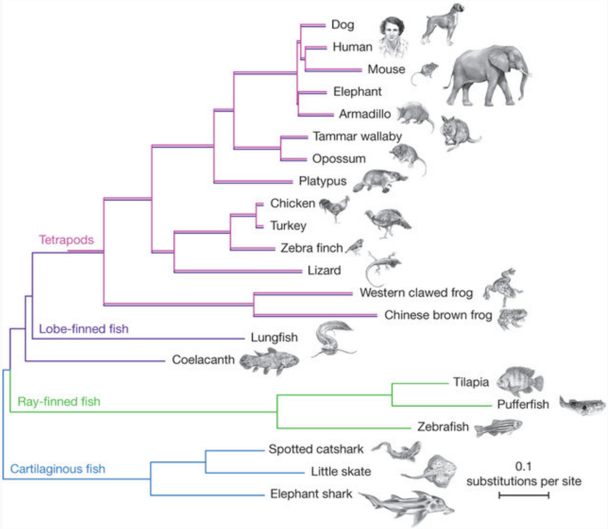
Generalizing Tree Probability Estimation via Bayesian Networks

Cheng Zhang and Frederick A. Matsen IV

December 1, 2018

Fred Hutchinson Cancer Research Center, Seattle, WA

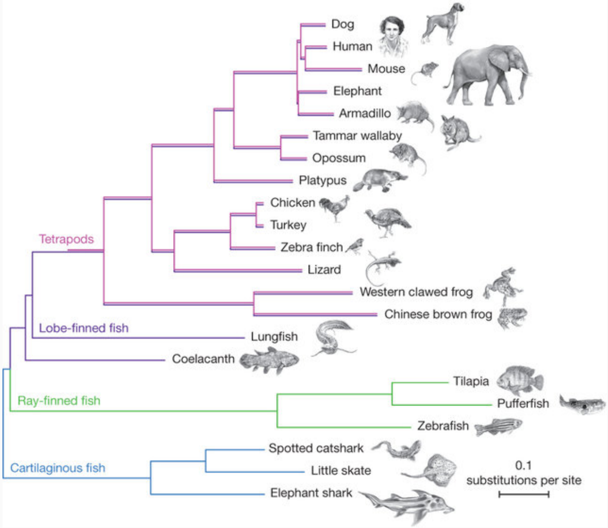
Phylogenetic Trees



Tree of life

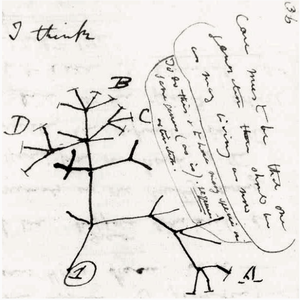
In **Molecular Evolution**, phylogenetic trees are used to model the evolutionary relationship among various biological species or other entities.

Phylogenetic Trees



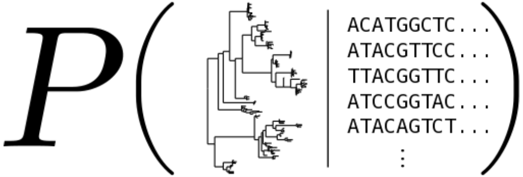
Tree of life

In **Molecular Evolution**, phylogenetic trees are used to model the evolutionary relationship among various biological species or other entities.


36
I think
B
C
D
1
A
Case must be that one generation then should be as many living as now.
To do this & to have many species in same genus (as is) requires extinction.

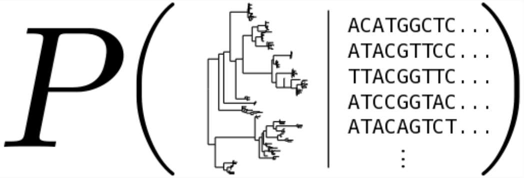
from Darwin's Notebook

Probability Estimation of Phylogenetic Trees



Markov chain Monte Carlo

Probability Estimation of Phylogenetic Trees

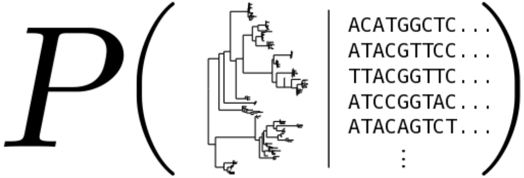


Markov chain Monte Carlo

Current approaches are unsatisfactory.

- Sample relative frequencies (**SRF**).
  - Do not generalize!
- Conditional clade distribution (**CCD**).
  - Not flexible enough for real data!

Probability Estimation of Phylogenetic Trees



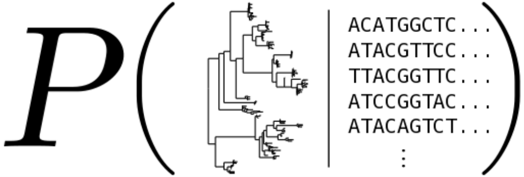
Markov chain Monte Carlo

Current approaches are unsatisfactory.

- Sample relative frequencies (**SRF**).
  - Do not generalize!
- Conditional clade distribution (**CCD**).
  - Not flexible enough for real data!

What is the best way to use MCMC samples?

Probability Estimation of Phylogenetic Trees



Markov chain Monte Carlo

Current approaches are unsatisfactory.

- Sample relative frequencies (**SRF**).
  - Do not generalize!
- Conditional clade distribution (**CCD**).
  - Not flexible enough for real data!

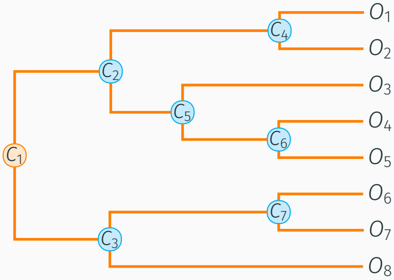
What is the best way to use MCMC samples?

Our Contribution: Subsplit Bayesian Networks. A general probability estimation framework for phylogenetic trees based on MCMC samples that

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

Key: harness the similarity of trees properly.

Problem Setup



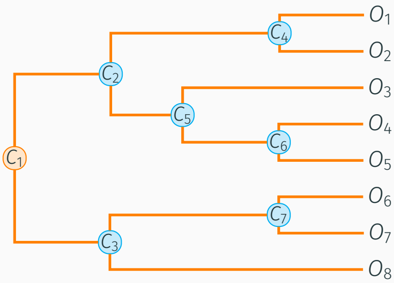
- Leaf label set $\mathcal{X} = \{O_1, \ldots, O_N\}$, each label represents a species.
- A *clade* X is a nonempty subset of $\mathcal{X}$.

$$C_5 = \{O_3, O_4, O_5\},\ C_7 = \{O_6, O_7\}.$$

- *Clade Decomposition*

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

Problem Setup



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

$$C_5 = \{O_3, O_4, O_5\}, \ C_7 = \{O_6, O_7\}.$$

- *Clade Decomposition*

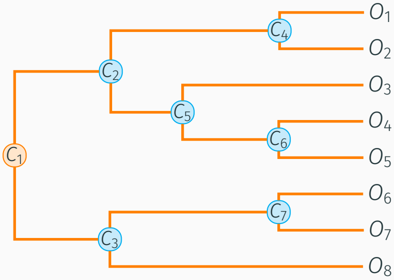
$$T_{\mathcal{C}} = \{C_2, C_3, C_4, C_5, C_6, C_7\}$$

A *subsplit* of a clade X is an ordered pair of disjoint subclades (Y, Z) such that $Y \cup Z = X, \ Y \succ Z$. Examples: $C_1 \to (C_2, C_3), \ C_2 \to (C_4, C_5)$.

Subsplit Decomposition

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

Problem Setup



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

$$C_5 = \{O_3, O_4, O_5\}, \; C_7 = \{O_6, O_7\}.$$

- *Clade Decomposition*

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

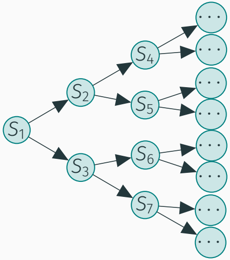
A *subsplit* of a clade X is an ordered pair of disjoint subclades (Y, Z) such that $Y \cup Z = X, \; Y \succ Z$. Examples: $C_1 \to (C_2, C_3), \; C_2 \to (C_4, C_5)$.

Subsplit Decomposition

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

$$p(T) = p(C_2, C_3)p(C_4, C_5|C_2, C_3)p(C_6|C_4, C_5)p(C_7|C_2, C_3)$$

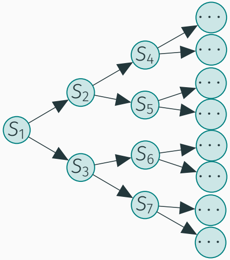
Subsplit Bayesian Networks

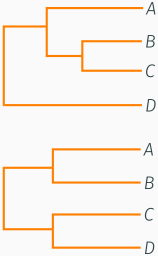


A **Subsplit Bayesian Network** on a leaf set $\mathcal{X}$ of size N is a Bayesian network

- nodes take on subsplit / singleton clade values.
- contains a full and complete binary tree.

Subsplit Bayesian Networks

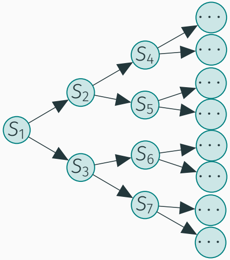


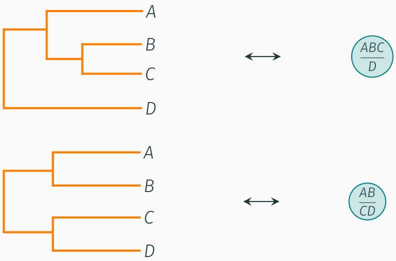


A **Subsplit Bayesian Network** on a leaf set $\mathcal{X}$ of size N is a Bayesian network

- nodes take on subsplit / singleton clade values.
- contains a full and complete binary tree.

Subsplit Bayesian Networks

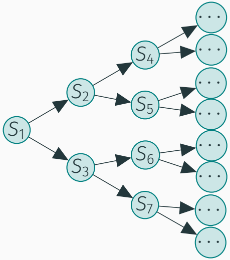


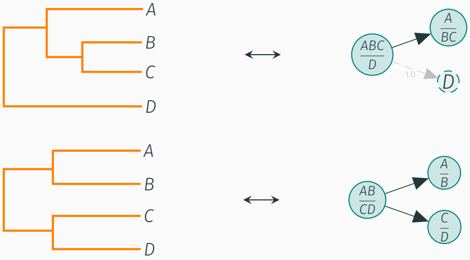


A **Subsplit Bayesian Network** on a leaf set $\mathcal{X}$ of size N is a Bayesian network

- nodes take on subsplit / singleton clade values.
- contains a full and complete binary tree.

Subsplit Bayesian Networks

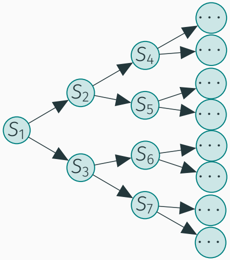


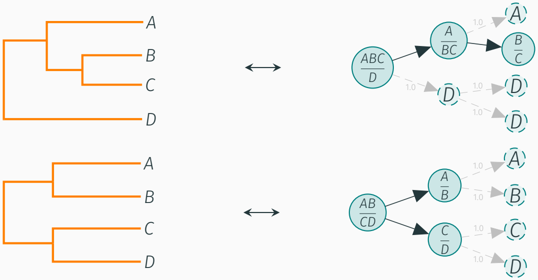


A **Subsplit Bayesian Network** on a leaf set $\mathcal{X}$ of size N is a Bayesian network

- nodes take on subsplit / singleton clade values.
- contains a full and complete binary tree.

Subsplit Bayesian Networks

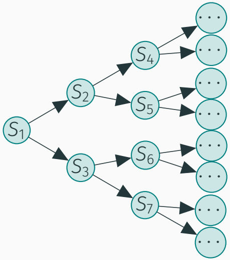


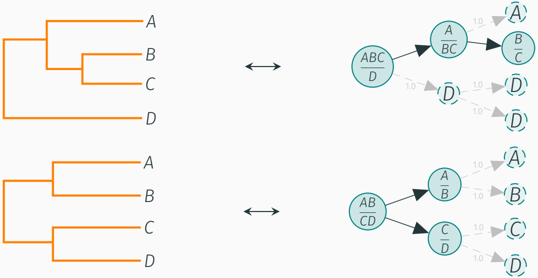


A **Subsplit Bayesian Network** on a leaf set $\mathcal{X}$ of size N is a Bayesian network

- nodes take on subsplit / singleton clade values.
- contains a full and complete binary tree.

Subsplit Bayesian Networks





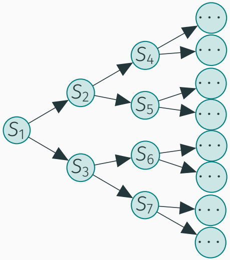
A **Subsplit Bayesian Network** on a leaf set $\mathcal{X}$ of size N is a Bayesian network

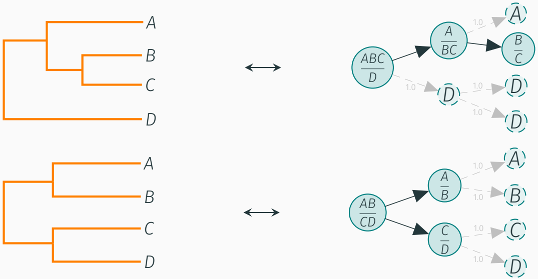
- nodes take on subsplit / singleton clade values.
- contains a full and complete binary tree.

SBN probability for rooted trees

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

Subsplit Bayesian Networks





A **Subsplit Bayesian Network** on a leaf set $\mathcal{X}$ of size N is a Bayesian network

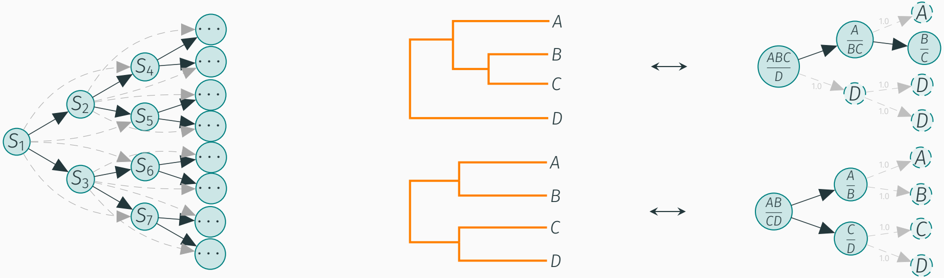
- nodes take on subsplit / singleton clade values.
- contains a full and complete binary tree.

SBN 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

Subsplit Bayesian Networks



A **Subsplit Bayesian Network** on a leaf set $\mathcal{X}$ of size N is a Bayesian network

- nodes take on subsplit / singleton clade values.
- contains a full and complete binary tree.

SBN 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 and are flexible.

Learning SBNs

Rooted Trees

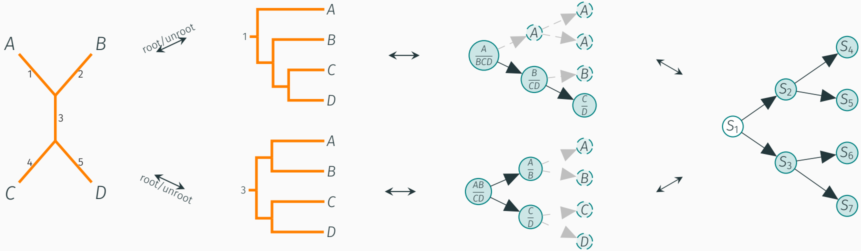
- maximum likelihood

Learning SBNs

Rooted Trees

- maximum likelihood

Unrooted Trees



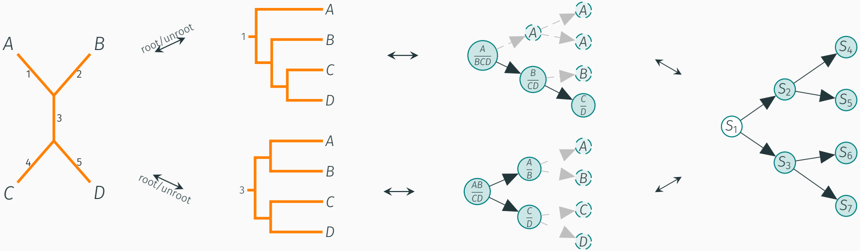
Learning SBNs

Rooted Trees

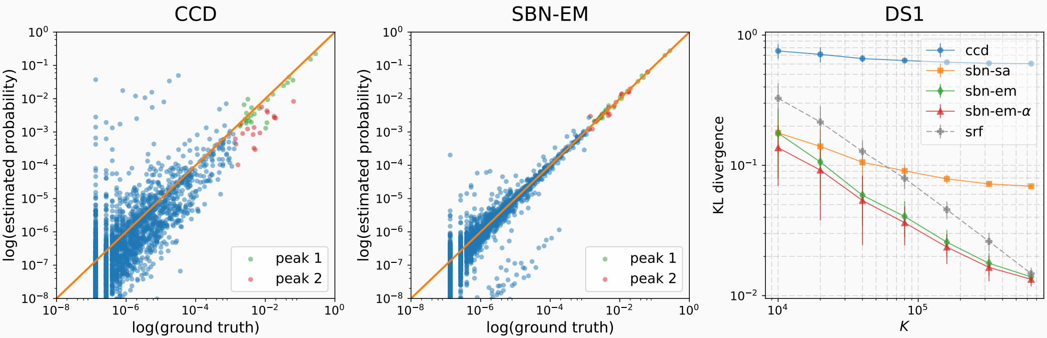
- maximum likelihood

Unrooted Trees

- Expectation Maximization
- simple averaging lower bound maximization
- incorporate regularization when necessary



Experiments



A real data set with multimodal posterior

Experiments

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**

Summary

Poster # 123

- We proposed a general framework for tree probability estimation based on **subsplit Bayesian networks**.
- SBNs exploit the similarity among trees to provide flexible probability estimators that generalize to unsampled trees.
- Future work
  - extends to general trees
  - structure learning of SBNs
  - deeper investigation on the effect of parameter sharing
  - applications in other probabilistic learning problems in tree spaces (e.g., MCMC transition kernel design and variational inference)