Improved Variational Bayesian Phylogenetic Inference with Normalizing Flows

Cheng Zhang

School of Mathematical Sciences & Center for Statistical Science

Peking University, Beijing, China

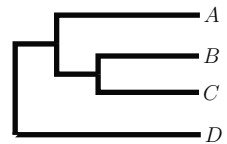
NeurIPS 2020

Phylogenetic Inference

Reconstruct the evolution history (i.e., *phylogenetic trees*) from molecular sequence data (e.g., DNA, RNA or protein sequences)

Taxa	Characters
Species A	ATGAACAT
Species B	ATGCACAC
Species C	ATGCATAT
Species D	ATGCATGC

Molecular Sequence Data



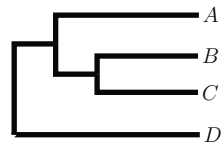
Phylogenetic Tree

Phylogenetic Inference

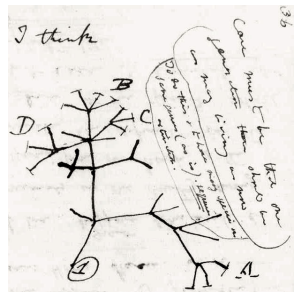
Reconstruct the evolution history (i.e., *phylogenetic trees*) from molecular sequence data (e.g., DNA, RNA or protein sequences)

Taxa	Characters
Species A	ATGAACAT
Species B	ATGCACAC
Species C	ATGCATAT
Species D	ATGCATGC

Molecular Sequence Data



Phylogenetic Tree

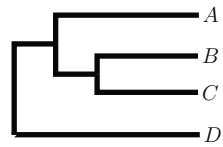


Phylogenetic Inference

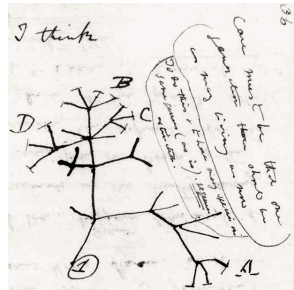
Reconstruct the evolution history (i.e., *phylogenetic trees*) from molecular sequence data (e.g., DNA, RNA or protein sequences)

Taxa	Characters
Species A	ATGAACAT
Species B	ATGCACAC
Species C	ATGCATAT
Species D	ATGCATGC

Molecular Sequence Data



Phylogenetic Tree



Lots of modern biological and medical applications!

BBC Sign in | Home | News | Sport | Reel | Worklife | T

NEWS

Home | US Election | Coronavirus | Video | World | Asia | UK | Business | Tech | Science | Stories

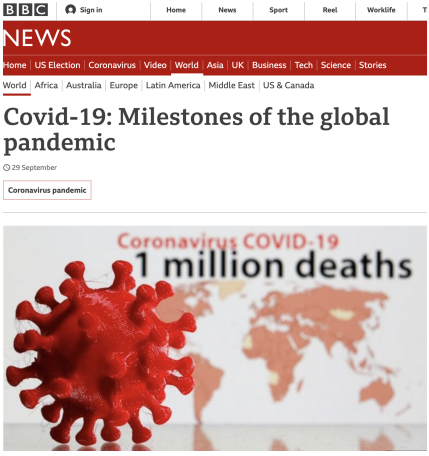
World | Africa | Australia | Europe | Latin America | Middle East | US & Canada

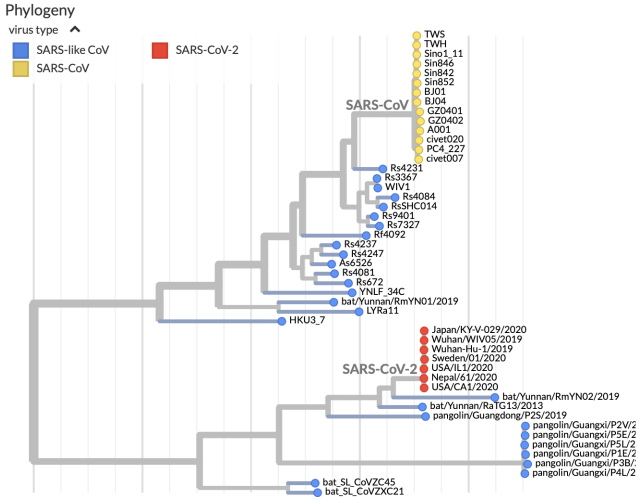
Covid-19: Milestones of the global pandemic

29 September

Coronavirus pandemic

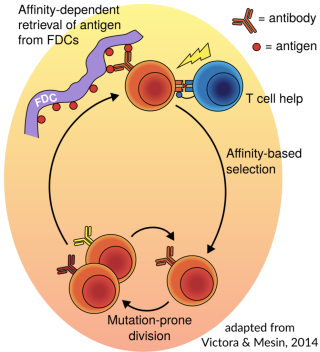
Genomic Epidemiology





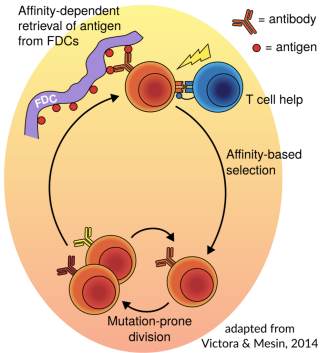
Immunology: B Cell Evolution

This happens inside of you!

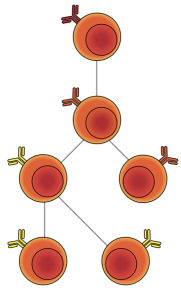


Immunology: B Cell Evolution

This happens inside of you!



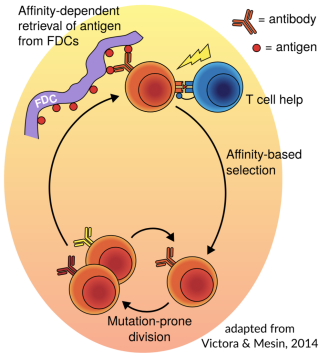
God sees

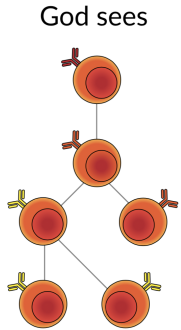


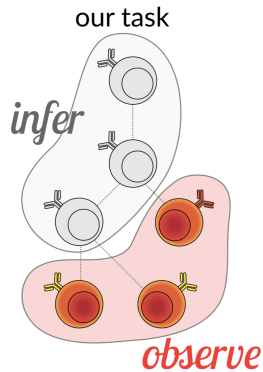
Immunology: B Cell Evolution

北京大学 PEKING UNIVERSITY

This happens inside of you!



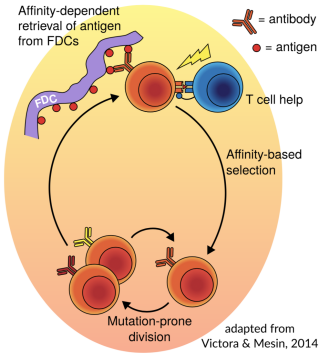


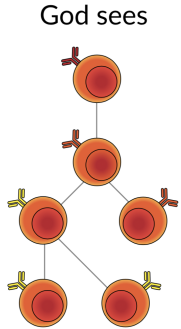


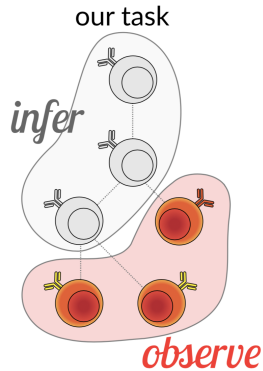
Immunology: B Cell Evolution

北京大学 PEKING UNIVERSITY

This happens inside of you!

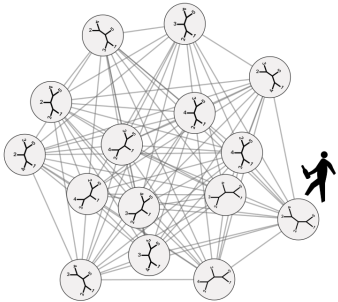




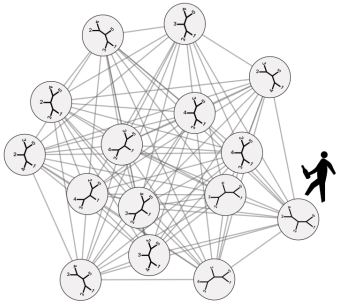


These inferences guide rational vaccine design.

Random-walk MCMC (MrBayes, BEAST)

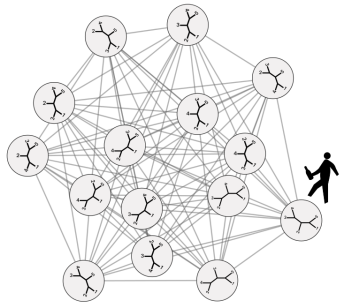


Random-walk MCMC (MrBayes, BEAST)



Challenges for MCMC

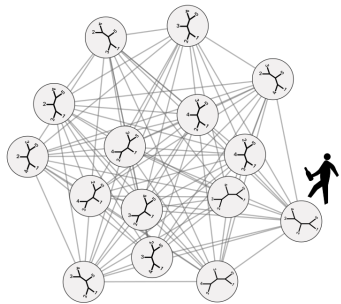
Random-walk MCMC (MrBayes, BEAST)



Challenges for MCMC

- Large search space: $(2n-5)!!$ unrooted trees (n taxa)

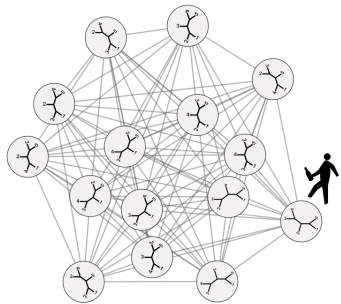
Random-walk MCMC (MrBayes, BEAST)



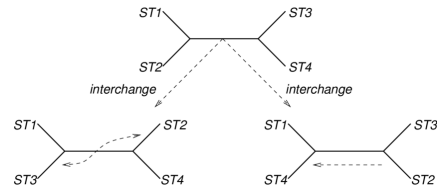
Challenges for MCMC

- Large search space: $(2n-5)!!$ unrooted trees (n taxa)
- Intertwined parameter space, low acceptance rate, hard to scale to data sets with many sequences.

Random-walk MCMC (MrBayes, BEAST)



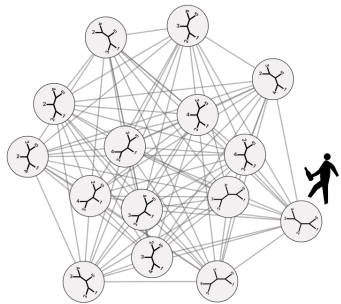
local tree transform

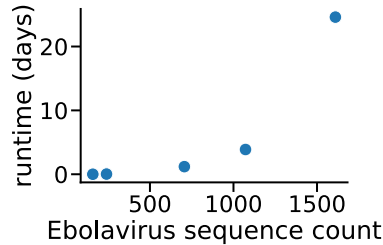


Challenges for MCMC

- Large search space: $(2n-5)!!$ unrooted trees (n taxa)
- Intertwined parameter space, low acceptance rate, hard to scale to data sets with many sequences.

Random-walk MCMC (MrBayes, BEAST)

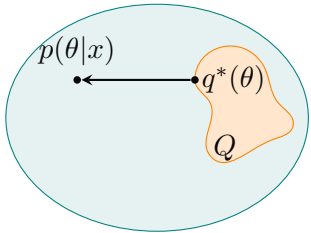




Challenges for MCMC

- Large search space: $(2n-5)!!$ unrooted trees (n taxa)
- Intertwined parameter space, low acceptance rate, hard to scale to data sets with many sequences.

Variational Inference



$$q^*(\theta) = \arg\min_{q \in Q} \mathrm{KL}\left(q(\theta) \| p(\theta|x)\right)$$

$$= \arg\max_{q \in Q} \mathbb{E}_{q(\theta)} \log \frac{p(x,\theta)}{q(\theta)}$$

- VI turns inference into optimization
- Specify a variational family of distributions over the model parameters

$$Q = \{q_\phi(\theta); \phi \in \Phi\}$$

- Fit the variational parameters ϕ to minimize the KL divergence, or equivalently, to maximize the evidence lower bound (ELBO)

Variational Bayesian Phylogenetic Inference

- Approximating Distribution:

$$Q_{\phi,\psi}(\tau, \boldsymbol{q}) \triangleq \underbrace{Q_{\phi}(\tau)}_{\text{tree topology}} \cdot \underbrace{Q_{\psi}(\boldsymbol{q}|\tau)}_{\text{branch length}}$$

[Zhang and Matsen IV, ICLR 2019]

Variational Bayesian Phylogenetic Inference

- Approximating Distribution:

$$Q_{\phi,\psi}(\tau, \boldsymbol{q}) \triangleq \underbrace{Q_{\phi}(\tau)}_{\text{tree topology}} \cdot \underbrace{Q_{\psi}(\boldsymbol{q}|\tau)}_{\text{branch length}}$$

- Multi-sample Lower Bound:

$$L^K(\phi, \psi) = \mathbb{E}_{Q_{\phi,\psi}(\tau^{1:K}, \boldsymbol{q}^{1:K})} \log \left(\frac{1}{K} \sum_{i=1}^{K} \frac{p(\boldsymbol{Y}|\tau^i, \boldsymbol{q}^i) p(\tau^i, \boldsymbol{q}^i)}{Q_{\phi}(\tau^i) Q_{\psi}(\boldsymbol{q}^i|\tau^i)} \right)$$

[Zhang and Matsen IV, ICLR 2019]

Variational Bayesian Phylogenetic Inference

- Approximating Distribution:

$$Q_{\phi,\psi}(\tau, \boldsymbol{q}) \triangleq \underbrace{Q_{\phi}(\tau)}_{\text{tree topology}} \cdot \underbrace{Q_{\psi}(\boldsymbol{q}|\tau)}_{\text{branch length}}$$

- Multi-sample Lower Bound:

$$L^K(\boldsymbol{\phi}, \boldsymbol{\psi}) = \mathbb{E}_{Q_{\boldsymbol{\phi},\boldsymbol{\psi}}(\tau^{1:K}, \boldsymbol{q}^{1:K})} \log \left(\frac{1}{K} \sum_{i=1}^{K} \frac{p(\boldsymbol{Y}|\tau^i, \boldsymbol{q}^i) p(\tau^i, \boldsymbol{q}^i)}{Q_{\boldsymbol{\phi}}(\tau^i) Q_{\boldsymbol{\psi}}(\boldsymbol{q}^i|\tau^i)} \right)$$

- Use **stochastic gradient ascent** (SGA) to maximize the lower bound:

$$\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\phi},\boldsymbol{\psi}} \; L^K(\boldsymbol{\phi}, \boldsymbol{\psi})$$

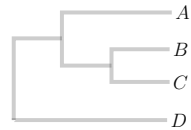
  Stochastic gradient estimators for the variational parameters

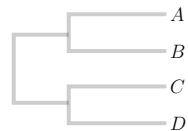
  ϕ : VIMCO/RWS, ψ : The Reparameterization Trick

[Zhang and Matsen IV, ICLR 2019]

Subsplit Bayesian Networks

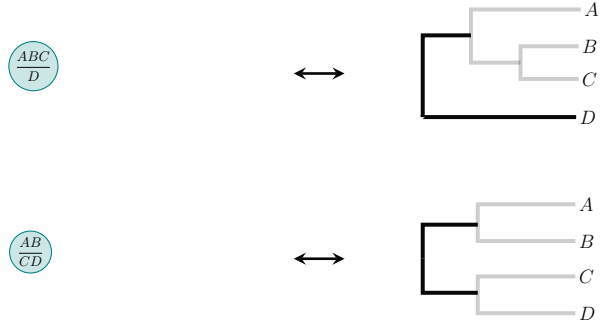
Inspired by previous works (Höhna and Drummond 2012, Larget 2013), we can decompose trees into local structures and encode the tree topology space via Bayesian networks! [Zhang and Matsen IV, NeurIPS 2018]





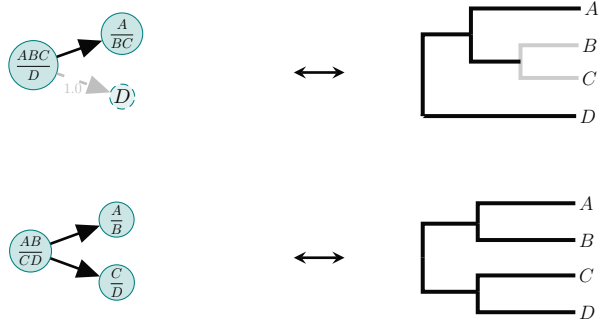
Subsplit Bayesian Networks

Inspired by previous works (Höhna and Drummond 2012, Larget 2013), we can decompose trees into local structures and encode the tree topology space via Bayesian networks! [Zhang and Matsen IV, NeurIPS 2018]



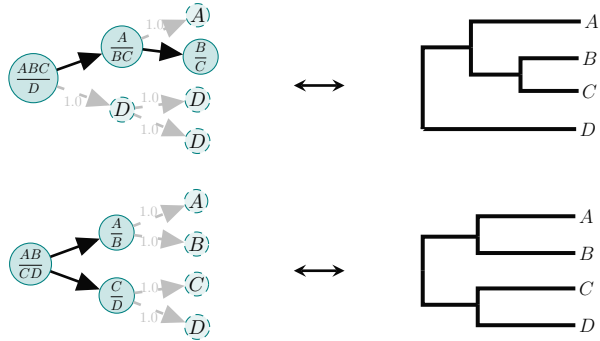
Subsplit Bayesian Networks

Inspired by previous works (Höhna and Drummond 2012, Larget 2013), we can decompose trees into local structures and encode the tree topology space via Bayesian networks! [Zhang and Matsen IV, NeurIPS 2018]



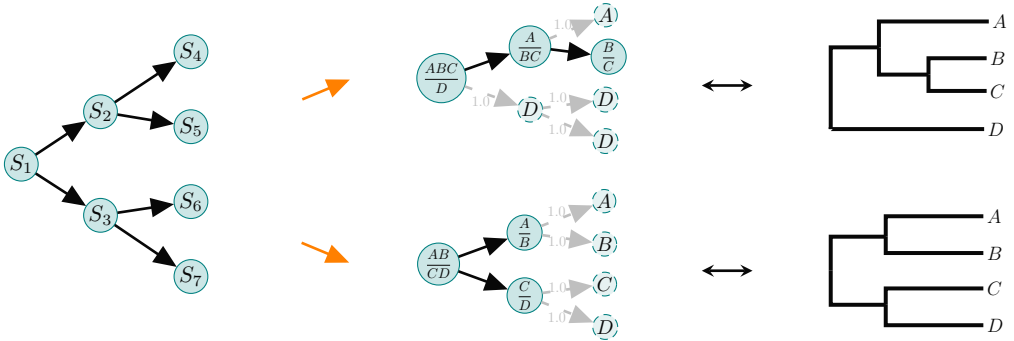
Subsplit Bayesian Networks

Inspired by previous works (Höhna and Drummond 2012, Larget 2013), we can decompose trees into local structures and encode the tree topology space via Bayesian networks! [Zhang and Matsen IV, NeurIPS 2018]

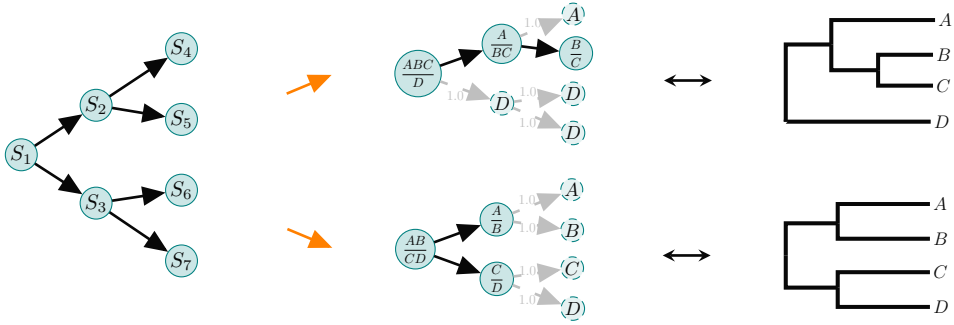


Subsplit Bayesian Networks

Inspired by previous works (Höhna and Drummond 2012, Larget 2013), we can decompose trees into local structures and encode the tree topology space via Bayesian networks! [Zhang and Matsen IV, NeurIPS 2018]



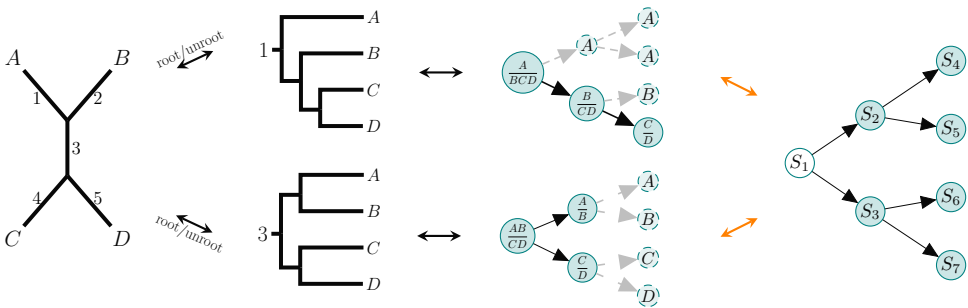
Tree Probability Estimation



Rooted Trees

$$p_{\text{sbn}}(T = \tau) = p(S_1 = s_1) \prod_{i>1} p(S_i = s_i | S_{\pi_i} = s_{\pi_i}).$$

Tree Probability Estimation



Unrooted Trees:

$$p_{\text{sbn}}(T^{\text{u}} = \tau) = \sum_{s_1 \sim \tau} p(S_1 = s_1) \prod_{i>1} p(S_i = s_i | S_{\pi_i} = s_{\pi_i}).$$

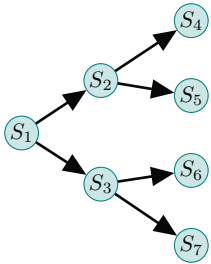
Tree Sampling

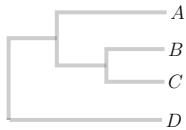
北京大学 PEKING UNIVERSITY

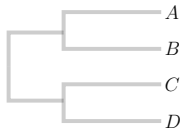
- Rooted Trees: ancestral sampling

Tree Sampling

- Rooted Trees: ancestral sampling

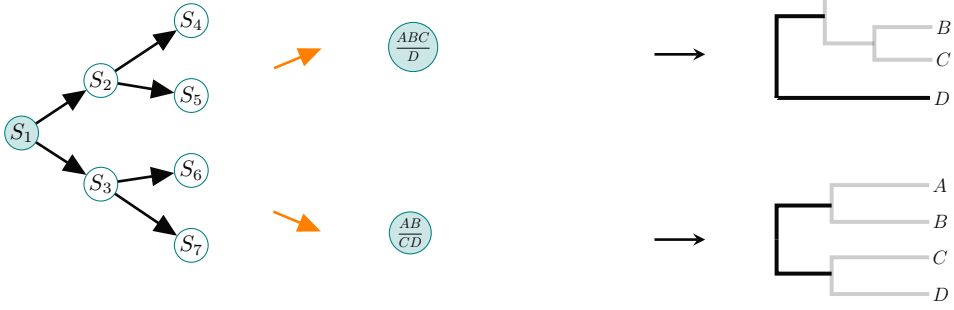






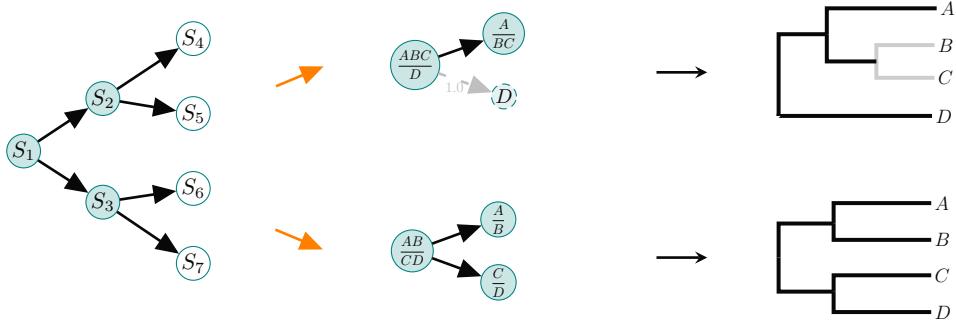
Tree Sampling

- Rooted Trees: ancestral sampling



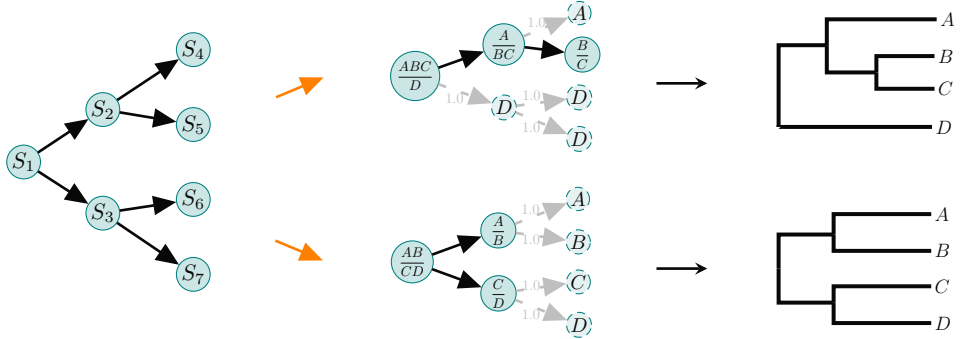
Tree Sampling

- Rooted Trees: ancestral sampling



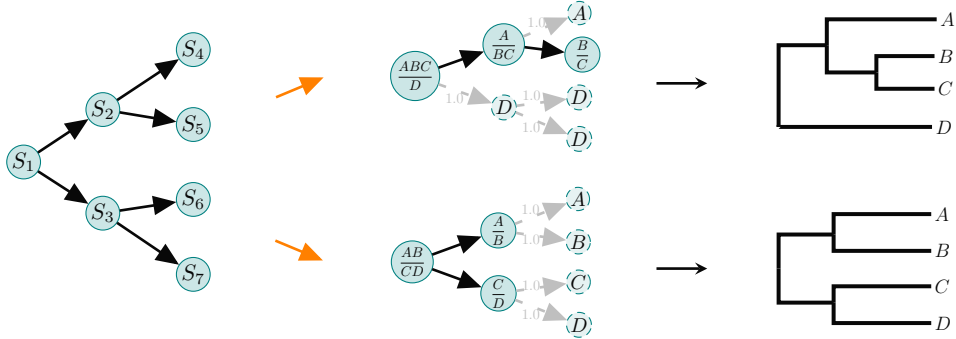
Tree Sampling

- Rooted Trees: ancestral sampling



Tree Sampling

- Rooted Trees: ancestral sampling



- Unrooted Trees: sample as rooted trees, then remove the roots

Structured Parameterization

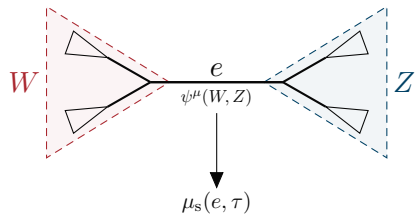
Branch length approximation

$$Q_{\boldsymbol{\psi}}(\boldsymbol{q}|\tau) = \prod_{e \in E(\tau)} p^{\text{Lognormal}}\left(q_e \mid \mu(e,\tau), \sigma(e,\tau)\right)$$

Amortization via local topological structures

- *Simple Split*

$$\mu_{\text{s}}(e,\tau) = \psi^{\mu}_{e/\tau},\ \sigma_{\text{s}}(e,\tau) = \psi^{\sigma}_{e/\tau}.$$



Structured Parameterization

Branch length approximation

$$Q_{\boldsymbol{\psi}}(\boldsymbol{q}|\tau) = \prod_{e \in E(\tau)} p^{\text{Lognormal}}\left(q_e \mid \mu(e,\tau), \sigma(e,\tau)\right)$$

Amortization via local topological structures

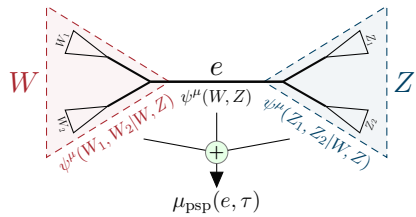
- *Simple Split*

$$\mu_{\text{s}}(e,\tau) = \psi^{\mu}_{e/\tau},\ \sigma_{\text{s}}(e,\tau) = \psi^{\sigma}_{e/\tau}.$$

- *Primary Subsplit Pair* (PSP)

$$\mu_{\text{psp}}(e,\tau) = \psi^{\mu}_{e/\tau} + \sum\nolimits_{s \in e/\!\!/\tau} \psi^{\mu}_{s}$$

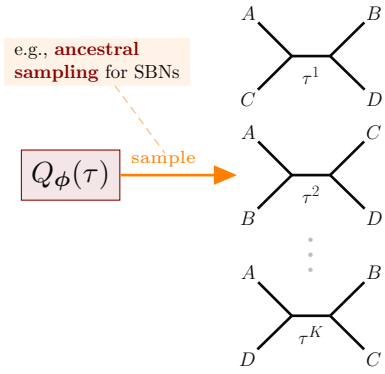
$$\sigma_{\text{psp}}(e,\tau) = \psi^{\sigma}_{e/\tau} + \sum\nolimits_{s \in e/\!\!/\tau} \psi^{\sigma}_{s}.$$



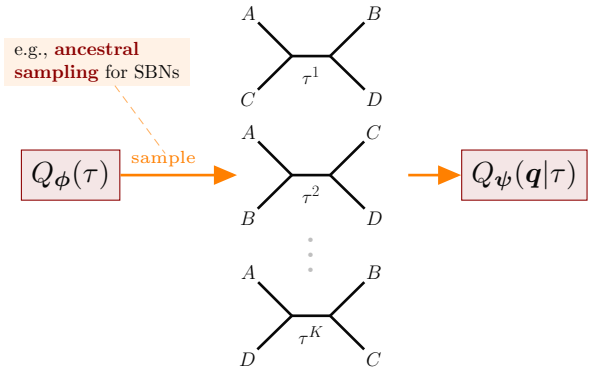
The VBPI Pipeline

$Q_\phi(\tau)$

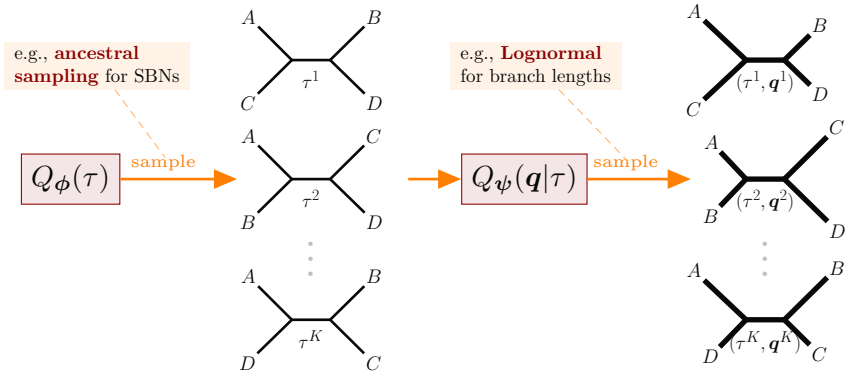
The VBPI Pipeline



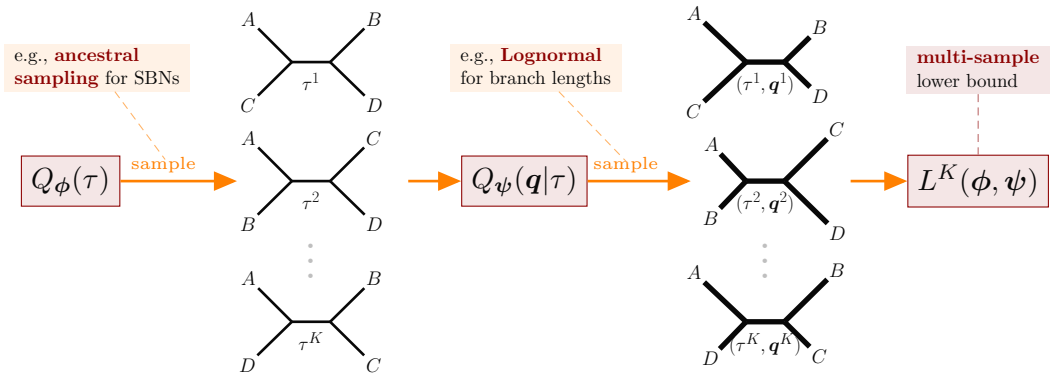
The VBPI Pipeline



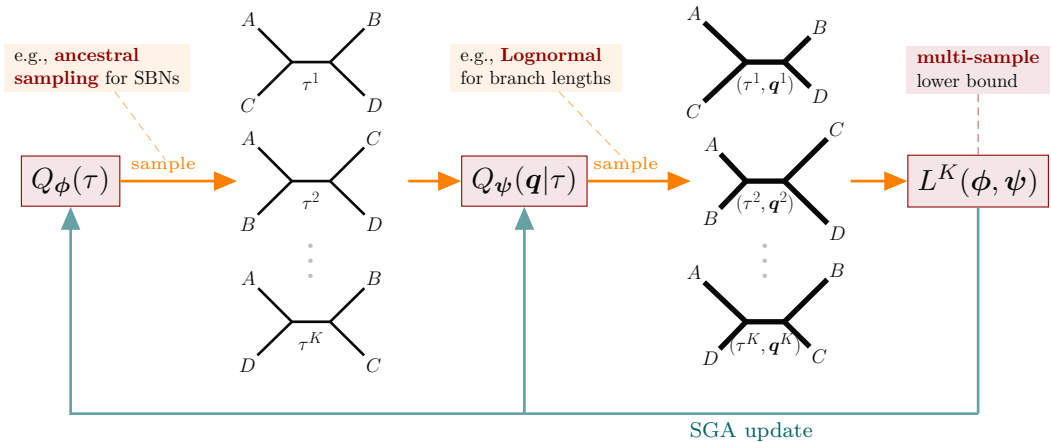
The VBPI Pipeline



The VBPI Pipeline



The VBPI Pipeline



Branch Length Approximation with Normalizing Flows

北京大学 PEKING UNIVERSITY

- Vanilla VBPI uses the diagonal Lognormal branch length approximation

$$Q_{\boldsymbol{\psi}}(\boldsymbol{q}|\tau) = \prod_{e \in E(\tau)} p^{\text{Lognormal}}\left(q_e \mid \mu(e,\tau), \sigma(e,\tau)\right)$$

Branch Length Approximation with Normalizing Flows

北京大学 PEKING UNIVERSITY

- Vanilla VBPI uses the diagonal Lognormal branch length approximation

$$Q_{\boldsymbol{\psi}}(\boldsymbol{q}|\tau) = \prod_{e \in E(\tau)} p^{\text{Lognormal}}\left(q_e \mid \mu(e,\tau), \sigma(e,\tau)\right)$$

- Improve the branch length approximation via normalizing flows?

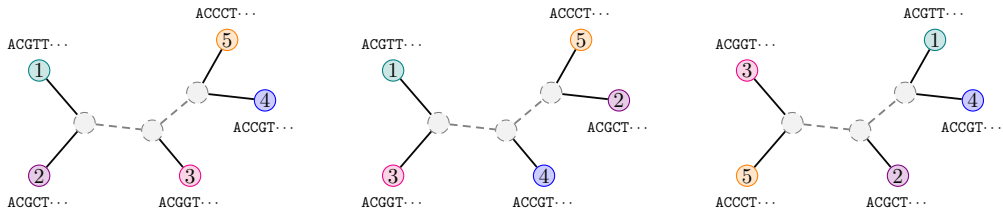
Branch Length Approximation with Normalizing Flows

北京大学 PEKING UNIVERSITY

- Vanilla VBPI uses the diagonal Lognormal branch length approximation

$$Q_{\psi}(\boldsymbol{q}|\tau) = \prod_{e \in E(\tau)} p^{\text{Lognormal}}\left(q_e \mid \mu(e,\tau), \sigma(e,\tau)\right)$$

- Improve the branch length approximation via normalizing flows?
- Hard to align the branch length vectors consistently across tree topologies!



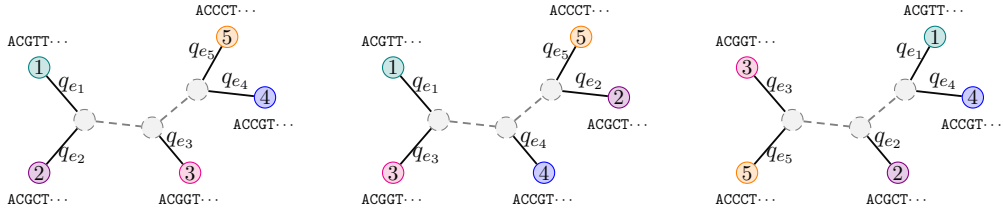
Branch Length Approximation with Normalizing Flows

北京大学 PEKING UNIVERSITY

- Vanilla VBPI uses the diagonal Lognormal branch length approximation

$$Q_{\psi}(\boldsymbol{q}|\tau) = \prod_{e \in E(\tau)} p^{\text{Lognormal}}\left(q_e \mid \mu(e,\tau), \sigma(e,\tau)\right)$$

- Improve the branch length approximation via normalizing flows?
- Hard to align the branch length vectors consistently across tree topologies!



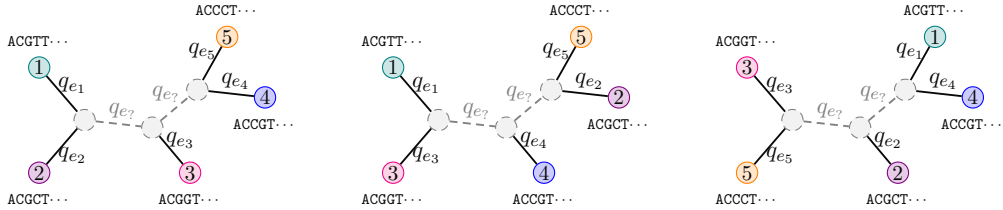
Branch Length Approximation with Normalizing Flows

北京大学 PEKING UNIVERSITY

- Vanilla VBPI uses the diagonal Lognormal branch length approximation

$$Q_{\psi}(\boldsymbol{q}|\tau) = \prod_{e \in E(\tau)} p^{\text{Lognormal}}\left(q_e \mid \mu(e,\tau), \sigma(e,\tau)\right)$$

- Improve the branch length approximation via normalizing flows?
- Hard to align the branch length vectors consistently across tree topologies!



Permutation Equivariant Planar Flows

- Standard planar transformation

$$z_i = x_i + \gamma_i a \left(\sum\nolimits_j w_j x_j + b \right), \quad i = 1, \ldots, d$$

- Structured planar transformation on phylogenetic trees

$$z_e = \tilde{q}_e + \gamma_e a \left(\sum\nolimits_{e' \in E(\tau)} w_{e'} \tilde{q}_{e'} + b \right), \quad \forall\, e \in E(\tau)$$

where

$$\gamma_e = \psi^{\gamma}_{e/\tau} + \sum_{s \in e /\!\!/ \tau} \psi^{\gamma}_s, \quad w_e = \psi^{w}_{e/\tau} + \sum_{s \in e /\!\!/ \tau} \psi^{w}_s$$

- The above planar transformation is permutation equivariant.

Permutation Equivariant RealNVP

- Standard affine coupling transformation

$$z_i = x_i,\ i \in S^c. \qquad z_i = x_i \exp(\alpha_i(\boldsymbol{x}_{S^c})) + \beta_i(\boldsymbol{x}_{S^c}),\ i \in S.$$

- Structured affine coupling transformation on phylogenetic trees

$$z_e = \tilde{q}_e,\ e \in S^c. \qquad z_e = \tilde{q}_e \exp\left(\alpha_e(\tilde{\boldsymbol{q}}_{S^c})\right) + \beta_e(\tilde{\boldsymbol{q}}_{S^c}),\ e \in S.$$

with a consistent pendant and interior bipartition $S \cup S^c$ of the edges across tree topologies, and permutation invariant α_e and β_e.

$$\begin{bmatrix} \alpha_e(\tilde{\boldsymbol{q}}_{S^c}) \\ \beta_e(\tilde{\boldsymbol{q}}_{S^c}) \end{bmatrix} = \begin{bmatrix} (\boldsymbol{w}_e^{\alpha})^T \\ \left(\boldsymbol{w}_e^{\beta}\right)^T \end{bmatrix} \rho\left(\sum_{e' \in S^c} \tilde{q}_{e'} \boldsymbol{w}_{e'} + \boldsymbol{b}\right) + \begin{bmatrix} b_e^{\alpha} \\ b_e^{\beta} \end{bmatrix}$$

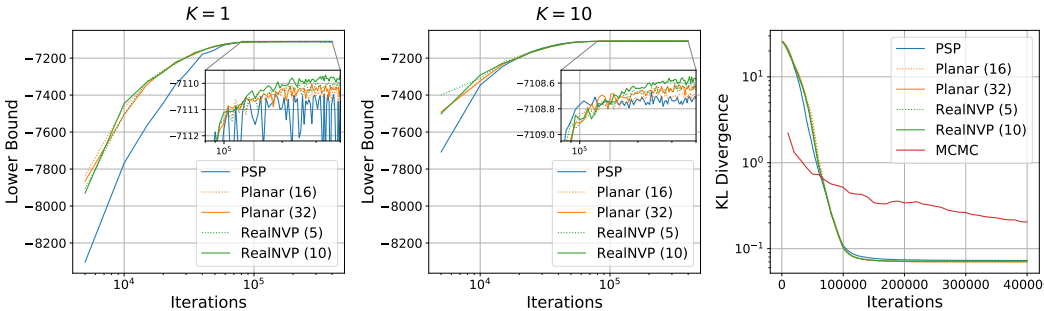
- The above affine coupling transformation is permutation equivariant.

Lower Bounds and Marginal Likelihoods

北京大学 PEKING UNIVERSITY

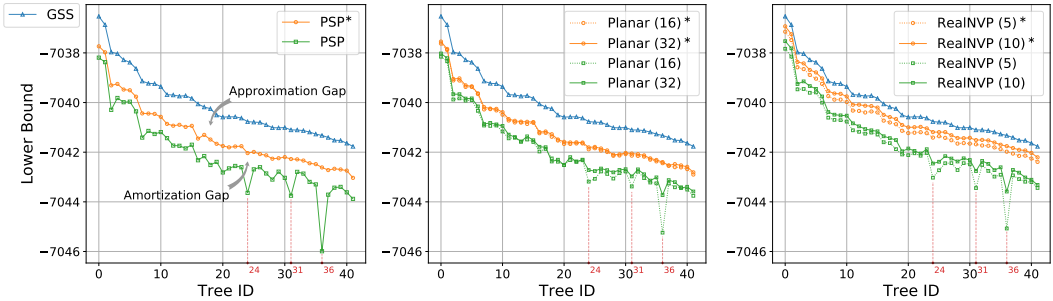
	Data set	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8
	# Taxa	27	29	36	41	50	50	59	64
	# Sites	1949	2520	1812	1137	378	1133	1824	1008
LB (K=1)	PSP	-7111.23(1.04)	-26369.63(0.69)	-33736.60(0.33)	-13332.37(0.54)	-8218.35(0.20)	-6729.27(0.50)	-37335.15(0.11)	-8655.48(0.38)
	Planar(16)	-7110.33(0.16)	-26368.80(0.27)	-33736.14(0.14)	-13331.92(0.11)	-8217.98(0.13)	-6728.89(0.18)	-37334.78(0.11)	-8655.15(0.17)
	Planar(32)	-7110.22(0.17)	-26368.69(0.23)	-33736.02(0.21)	-13331.73(0.12)	-8217.90(0.14)	-6728.68(0.19)	-37334.60(0.12)	-8654.97(0.16)
	RealNVP(5)	-7110.12(0.13)	-26368.75(0.24)	-33735.86(0.10)	-13331.71(0.11)	-8217.80(0.14)	-6728.54(0.15)	-37334.44(0.11)	-8654.62(0.13)
	RealNVP(10)	**-7109.80(0.11)**	**-26368.59(0.23)**	**-33735.81(0.12)**	**-13331.39(0.08)**	**-8217.56(0.12)**	**-6728.04(0.14)**	**-37333.94(0.09)**	**-8654.02(0.12)**
LB (K=10)	PSP	-7108.73(0.02)	-26367.88(0.02)	-33735.29(0.02)	-13330.34(0.03)	-8215.57(0.04)	-6725.48(0.04)	-37332.69(0.03)	-8651.88(0.04)
	Planar(16)	-7108.70(0.02)	-26367.80(0.01)	-33735.21(0.01)	-13330.28(0.02)	-8215.44(0.04)	-6725.42(0.04)	-37332.50(0.03)	-8651.80(0.04)
	Planar(32)	-7108.64(0.02)	-26367.77(0.01)	-33735.17(0.01)	-13330.22(0.02)	-8215.37(0.03)	-6725.32(0.04)	-37332.43(0.03)	-8651.72(0.04)
	RealNVP(5)	-7108.63(0.02)	-26367.77(0.01)	-33735.18(0.01)	-13330.22(0.02)	-8215.36(0.03)	-6725.33(0.04)	-37332.42(0.03)	-8651.62(0.04)
	RealNVP(10)	**-7108.58(0.02)**	**-26367.75(0.01)**	**-33735.16(0.01)**	**-13330.16(0.02)**	**-8215.29(0.03)**	**-6725.18(0.04)**	**-37332.30(0.02)**	**-8651.41(0.03)**
ML	PSP	-7108.39(0.18)	-26367.71(0.08)	-33735.09(0.10)	-13329.93(0.21)	-8214.44(0.48)	-6724.13(0.48)	-37331.92(0.32)	-8650.12(0.58)
	Planar(16)	-7108.39(0.15)	-26367.70(0.07)	-33735.09(0.07)	-13329.93(0.17)	-8214.49(0.42)	-6724.25(0.45)	-37331.91(0.26)	-8650.42(0.52)
	Planar(32)	-7108.40(0.14)	-26367.70(0.06)	**-33735.09(0.05)**	-13329.93(0.16)	-8214.50(0.38)	-6724.19(0.44)	-37331.93(0.23)	-8650.40(0.50)
	RealNVP(5)	-7108.40(0.14)	**-26367.71(0.04)**	-33735.09(0.06)	-13329.92(0.16)	-8214.50(0.38)	-6724.28(0.39)	**-37331.92(0.22)**	-8650.46(0.44)
	RealNVP(10)	**-7108.39(0.11)**	**-26367.71(0.04)**	**-33735.09(0.05)**	**-13329.92(0.13)**	-8214.51(0.36)	**-6724.25(0.37)**	**-37331.90(0.22)**	**-8650.42(0.41)**
	SS	-7108.42(0.18)	-26367.57(0.48)	-33735.44(0.50)	-13330.06(0.54)	**-8214.51(0.28)**	-6724.07(0.86)	-37332.76(2.42)	-8649.88(1.75)

Computational Complexity and Convergence



- Achieve comparable approximation quality when PSP converges, and quickly surpass PSP as the number of iterations increases.
- Maintain the speed advantage of PSP when compared to MCMC.

Approximation and Amortization Gaps



Gap	PSP		Planar (16)		Planar (32)		RealNVP (5)		RealNVP (10)	
	Tree 36	All	Tree 36	All	Tree 36	All	Tree 36	All	Tree 36	All
Approximation	1.29	1.21	1.12	1.08	1.07	1.02	0.65	0.62	**0.43**	**0.40**
Amortization	3.37	0.84	2.80	0.82	**1.33**	**0.72**	3.10	0.98	1.83	0.93
Inference	4.66	2.05	3.92	1.90	2.40	1.74	3.75	1.60	**2.26**	**1.33**

Thank you!