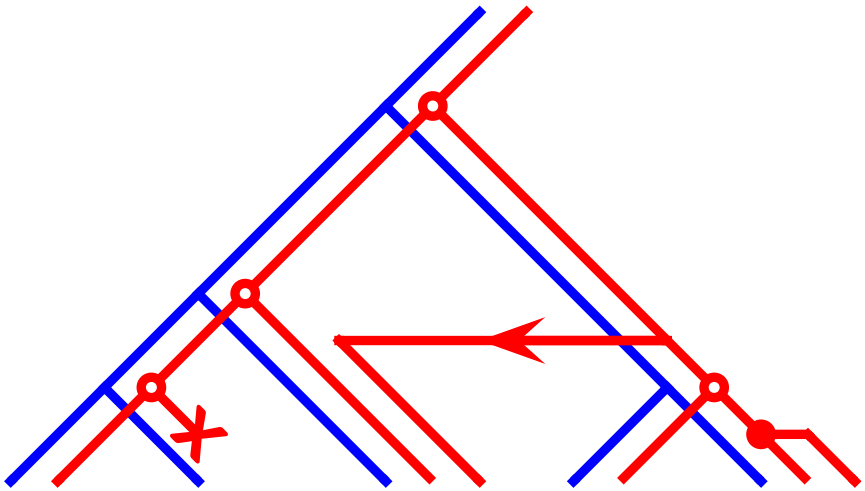
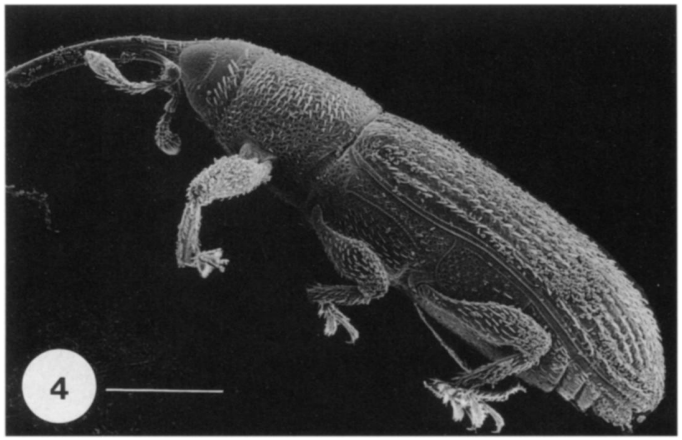
Bayesian Reconstruction of Coevolutionary Histories

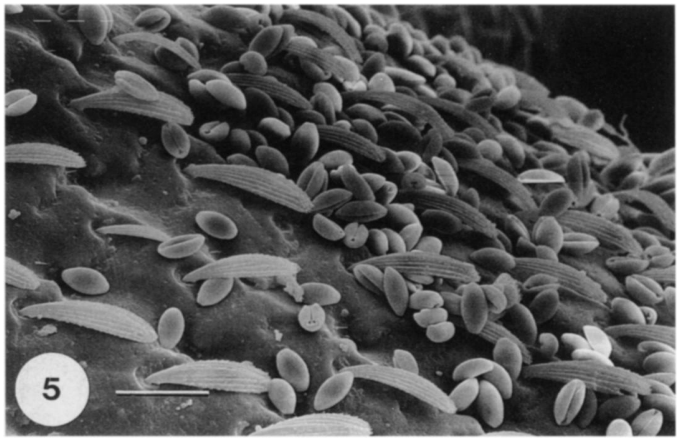
Arman Bilge

March 13, 2013



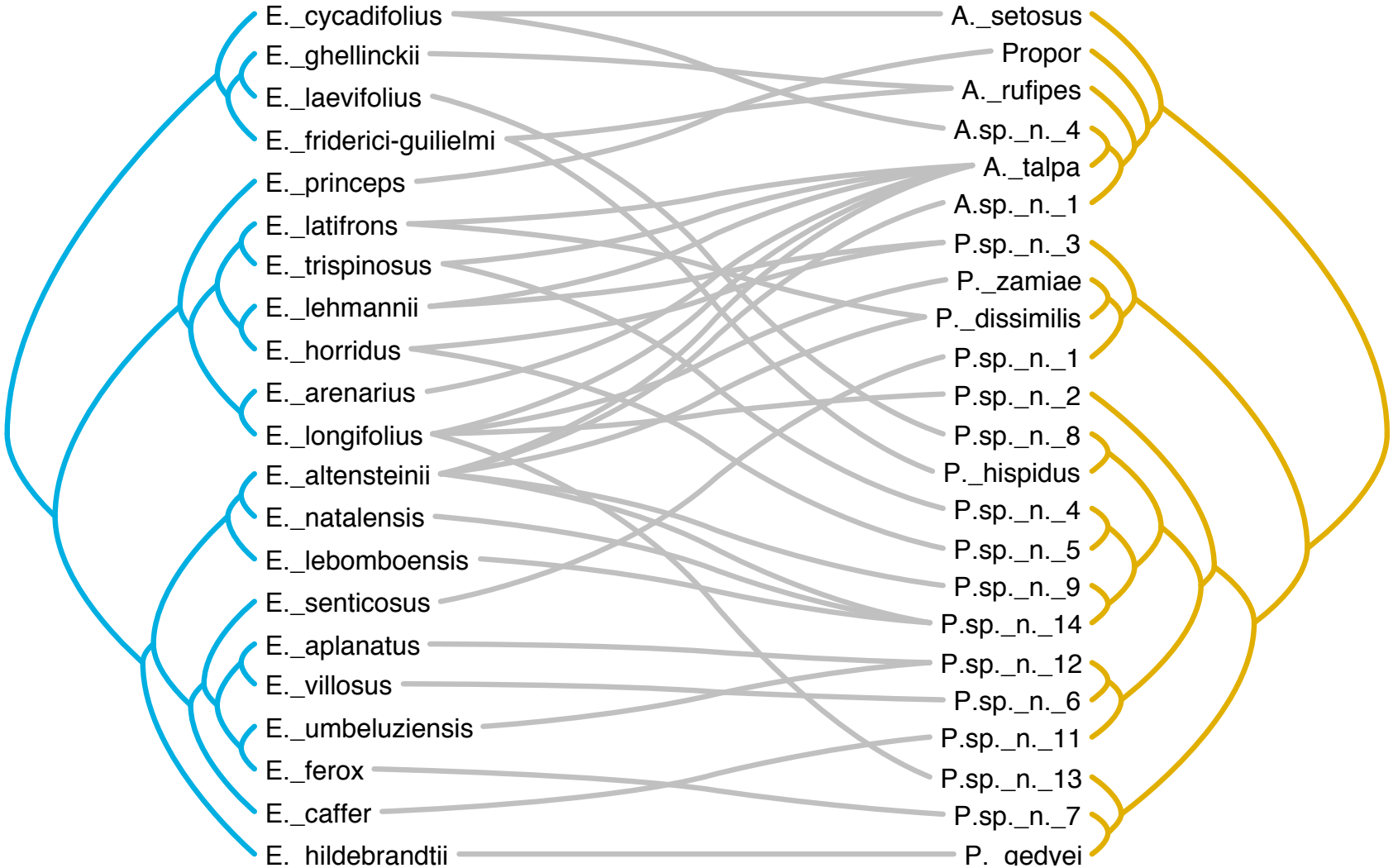
Phylogenetic Evidence for Cycad–Weevil Coevolution





How should we compare the phylogenies of symbionts?

How should we compare the phylogenies of symbionts?



The Cophylogeny Reconciliation Problem

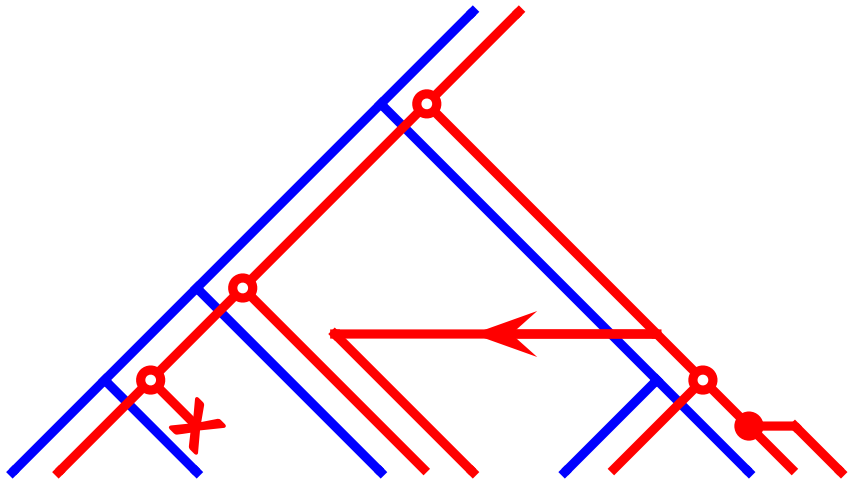
- What were the ancestral host–symbiont partnerships?

The Cophylogeny Reconciliation Problem

- What were the ancestral host–symbiont partnerships?
- Like ancestral state reconstruction, except with prior on rate matrix based on host phylogeny

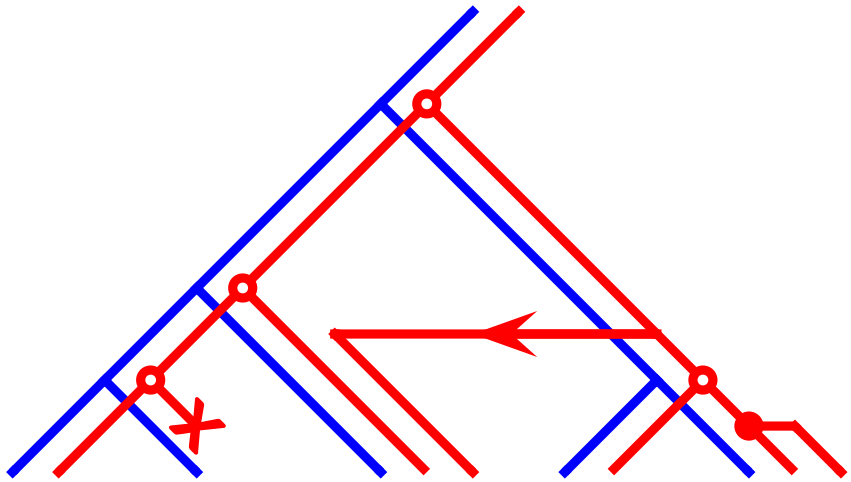
The Cophylogeny Reconciliation Problem

- What were the ancestral host–symbiont partnerships?
- Like ancestral state reconstruction, except with prior on rate matrix based on host phylogeny
- Discrepancies between host and symbiont phylogenies explained by coevolutionary events
  - Cospeciation, duplication, host-switch, and loss



The Cophylogeny Reconciliation Problem

- What were the ancestral host–symbiont partnerships?
- Like ancestral state reconstruction, except with prior on rate matrix based on host phylogeny
- Discrepancies between host and symbiont phylogenies explained by coevolutionary events
  - Cospeciation, duplication, host-switch, and loss
- Also applies to gene tree–species tree reconciliation, particularly involving horizontal transfer events



Existing Work on the Cophylogeny Reconciliation Problem

Parsimony Methods

- Tarzan, CoRe-PA, Jane, TreeMap, AnGST, RANGER-DTL

Existing Work on the Cophylogeny Reconciliation Problem

Parsimony Methods

- Tarzan, CoRe-PA, Jane, TreeMap, AnGST, RANGER-DTL
- Both trees must be fixed—no accommodation for uncertainty

Existing Work on the Cophylogeny Reconciliation Problem

Parsimony Methods

- Tarzan, CoRe-PA, Jane, TreeMap, AnGST, RANGER-DTL
- Both trees must be fixed—no accommodation for uncertainty
- How to assign penalties to events or weigh numerous equally parsimonious solutions?

Existing Work on the Cophylogeny Reconciliation Problem

Parsimony Methods

- Tarzan, CoRe-PA, Jane, TreeMap, AnGST, RANGER-DTL
- Both trees must be fixed—no accommodation for uncertainty
- How to assign penalties to events or weigh numerous equally parsimonious solutions?
- Primitive consideration of node timing info

Existing Work on the Cophylogeny Reconciliation Problem

Parsimony Methods

- Tarzan, CoRe-PA, Jane, TreeMap, AnGST, RANGER-DTL
- Both trees must be fixed—no accommodation for uncertainty
- How to assign penalties to events or weigh numerous equally parsimonious solutions?
- Primitive consideration of node timing info
- No consideration of geographical data

Existing Work on the Cophylogeny Reconciliation Problem

Parsimony Methods

- Tarzan, CoRe-PA, Jane, TreeMap, AnGST, RANGER-DTL
- Both trees must be fixed—no accommodation for uncertainty
- How to assign penalties to events or weigh numerous equally parsimonious solutions?
- Primitive consideration of node timing info
- No consideration of geographical data

Probabilistic Methods

- Huelsenbeck et al. (2000), Charleston (2009), Faria et al. (2013), JPrIME DLTRS model

Existing Work on the Cophylogeny Reconciliation Problem

Parsimony Methods

- Tarzan, CoRe-PA, Jane, TreeMap, AnGST, RANGER-DTL
- Both trees must be fixed—no accommodation for uncertainty
- How to assign penalties to events or weigh numerous equally parsimonious solutions?
- Primitive consideration of node timing info
- No consideration of geographical data

Probabilistic Methods

- Huelsenbeck et al. (2000), Charleston (2009), Faria et al. (2013), JPrIME DLTRS model
- Each relies on various assumptions and simplifications

A Bayesian Interpretation of Cophylogeny

$$\overbrace{P\left(H, S, \mathcal{R} \mid D\right)}^{\text{probability of reconstruction}}$$

H = host tree, S = symbiont tree, $\mathcal{R}$ = reconciliation,
θ = model parameters, $D = (d_H, d_S)$ = sequence data

A Bayesian Interpretation of Cophylogeny

$$\overbrace{P\left(H, S, \mathcal{R} \mid D\right)}^{\text{probability of reconstruction}} \propto \int_{\theta} \overbrace{P\left(d_H, d_S \mid H, S, \mathcal{R}, \theta\right)}^{\text{likelihood}} \overbrace{P\left(H, S, \mathcal{R}, \theta\right)}^{\text{prior}} \, \mathrm{d}\theta$$

H = host tree, S = symbiont tree, $\mathcal{R}$ = reconciliation,
θ = model parameters, $D = (d_H, d_S)$ = sequence data

A Bayesian Interpretation of Cophylogeny

$$\overbrace{P\left(H, S, \mathcal{R} \mid D\right)}^{\text{probability of reconstruction}} \propto \int_{\theta} \overbrace{P\left(d_H, d_S \mid H, S, \mathcal{R}, \theta\right)}^{\text{likelihood}} \overbrace{P\left(H, S, \mathcal{R}, \theta\right)}^{\text{prior}} \, \mathrm{d}\theta$$

Likelihood

$$P\left(d_H, d_S \mid H, S, \mathcal{R}, \theta\right) = P\left(d_H \mid H, S, \mathcal{R}, \theta\right) P\left(d_S \mid H, S, \mathcal{R}, \theta\right)$$

$H =$ host tree, $S =$ symbiont tree, $\mathcal{R} =$ reconciliation,
$\theta =$ model parameters, $D = (d_H, d_S) =$ sequence data

A Bayesian Interpretation of Cophylogeny

$$\overbrace{P\left(H, S, \mathcal{R} \mid D\right)}^{\text{probability of reconstruction}} \propto \int_{\theta} \overbrace{P\left(d_H, d_S \mid H, S, \mathcal{R}, \theta\right)}^{\text{likelihood}} \overbrace{P\left(H, S, \mathcal{R}, \theta\right)}^{\text{prior}} \, \mathrm{d}\theta$$

Likelihood

$$\begin{aligned} P\left(d_H, d_S \mid H, S, \mathcal{R}, \theta\right) &= P\left(d_H \mid H, S, \mathcal{R}, \theta\right) P\left(d_S \mid H, S, \mathcal{R}, \theta\right) \\ &= \underbrace{P\left(d_H \mid H\right)}_{\text{tree likelihood}} \underbrace{P\left(d_S \mid S\right)}_{\text{tree likelihood}} \end{aligned}$$

$H =$ host tree, $S =$ symbiont tree, $\mathcal{R} =$ reconciliation,
$\theta =$ model parameters, $D = (d_H, d_S) =$ sequence data

A Bayesian Interpretation of Cophylogeny

$$\overbrace{P\left(H, S, \mathcal{R} \mid D\right)}^{\text{probability of reconstruction}} \propto \int_{\theta} \overbrace{P\left(d_H, d_S \mid H, S, \mathcal{R}, \theta\right)}^{\text{likelihood}} \overbrace{P\left(H, S, \mathcal{R}, \theta\right)}^{\text{prior}} \, \mathrm{d}\theta$$

Prior

$$P\left(H, S, \mathcal{R}, \theta\right) = P\left(S \mid H, \mathcal{R}, \theta\right) P\left(H, \mathcal{R}, \theta\right)$$

$H =$ host tree, $S =$ symbiont tree, $\mathcal{R} =$ reconciliation,
$\theta =$ model parameters, $D = (d_H, d_S) =$ sequence data

A Bayesian Interpretation of Cophylogeny

$$\overbrace{P\left(H, S, \mathcal{R} \mid D\right)}^{\text{probability of reconstruction}} \propto \int_{\theta} \overbrace{P\left(d_H, d_S \mid H, S, \mathcal{R}, \theta\right)}^{\text{likelihood}} \overbrace{P\left(H, S, \mathcal{R}, \theta\right)}^{\text{prior}} \, \mathrm{d}\theta$$

Prior

$$\begin{aligned} P\left(H, S, \mathcal{R}, \theta\right) &= P\left(S \mid H, \mathcal{R}, \theta\right) P\left(H, \mathcal{R}, \theta\right) \\ &= \underbrace{P\left(S \mid H, \mathcal{R}, \theta\right)}_{?} \underbrace{P\left(H\right) P\left(\mathcal{R}\right) P\left(\theta\right)}_{\text{existing/trivial priors}} \end{aligned}$$

$H =$ host tree, $S =$ symbiont tree, $\mathcal{R} =$ reconciliation,
$\theta =$ model parameters, $D = (d_H, d_S) =$ sequence data

Computing $P\left(S \mid H, \mathcal{R}, \theta\right)$

- There are infinite histories that may yield S under H and $\mathcal{R}$
  - Particularly confounding due to loss events
  - Probability cannot be integrated analytically

Computing $P\left(S \mid H, \mathcal{R}, \theta\right)$

- There are infinite histories that may yield S under H and $\mathcal{R}$
  - Particularly confounding due to loss events
  - Probability cannot be integrated analytically
- As a simplification, I consider only "observable" events

Computing $P\left(S \mid H, \mathcal{R}, \theta\right)$

- There are infinite histories that may yield S under H and $\mathcal{R}$
  - Particularly confounding due to loss events
  - Probability cannot be integrated analytically
- As a simplification, I consider only "observable" events
- Assume that symbiont always cospeciates with its host

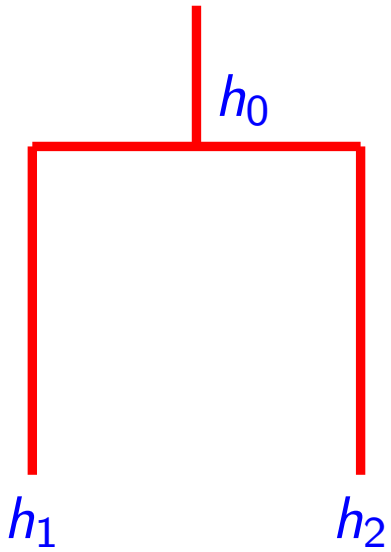
Computing $P\left(S \mid H, \mathcal{R}, \theta\right)$

- There are infinite histories that may yield S under H and $\mathcal{R}$
  - Particularly confounding due to loss events
  - Probability cannot be integrated analytically
- As a simplification, I consider only "observable" events
- Assume that symbiont always cospeciates with its host
- Other events independent of host and modelled as Poisson processes with independent rates

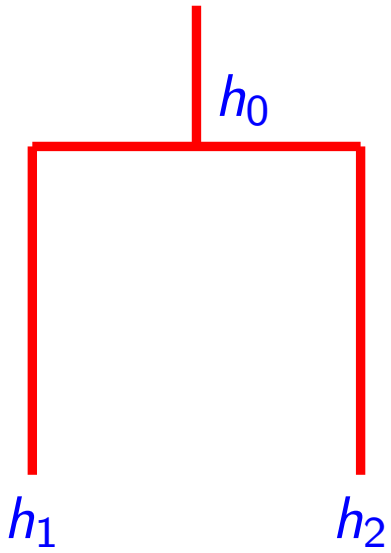
Computing $P\left(S \mid H, \mathcal{R}, \theta\right)$

- There are infinite histories that may yield S under H and $\mathcal{R}$
  - Particularly confounding due to loss events
  - Probability cannot be integrated analytically
- As a simplification, I consider only "observable" events
- Assume that symbiont always cospeciates with its host
- Other events independent of host and modelled as Poisson processes with independent rates
- Symbiont and host must be contemporaneous

Computing $P\left(S \mid H, \mathcal{R}, \theta\right)$

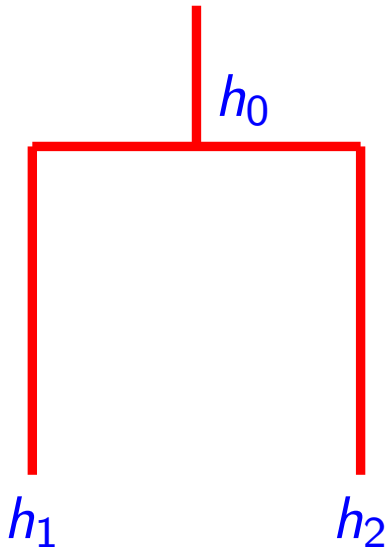


Computing $P\left(S \mid H, \mathcal{R}, \theta\right)$



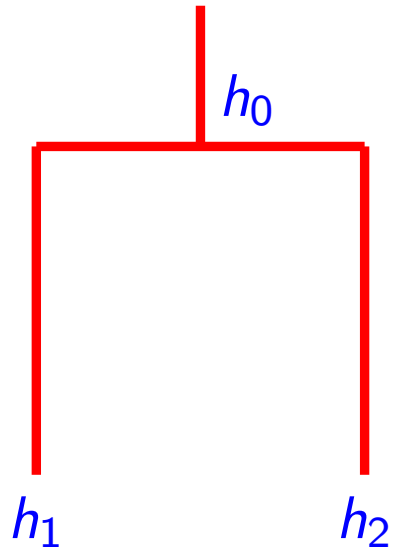
- What is the relationship of h_0 to h_1 and h_2 in the host tree?

Computing $P\left(S \mid H, \mathcal{R}, \theta\right)$



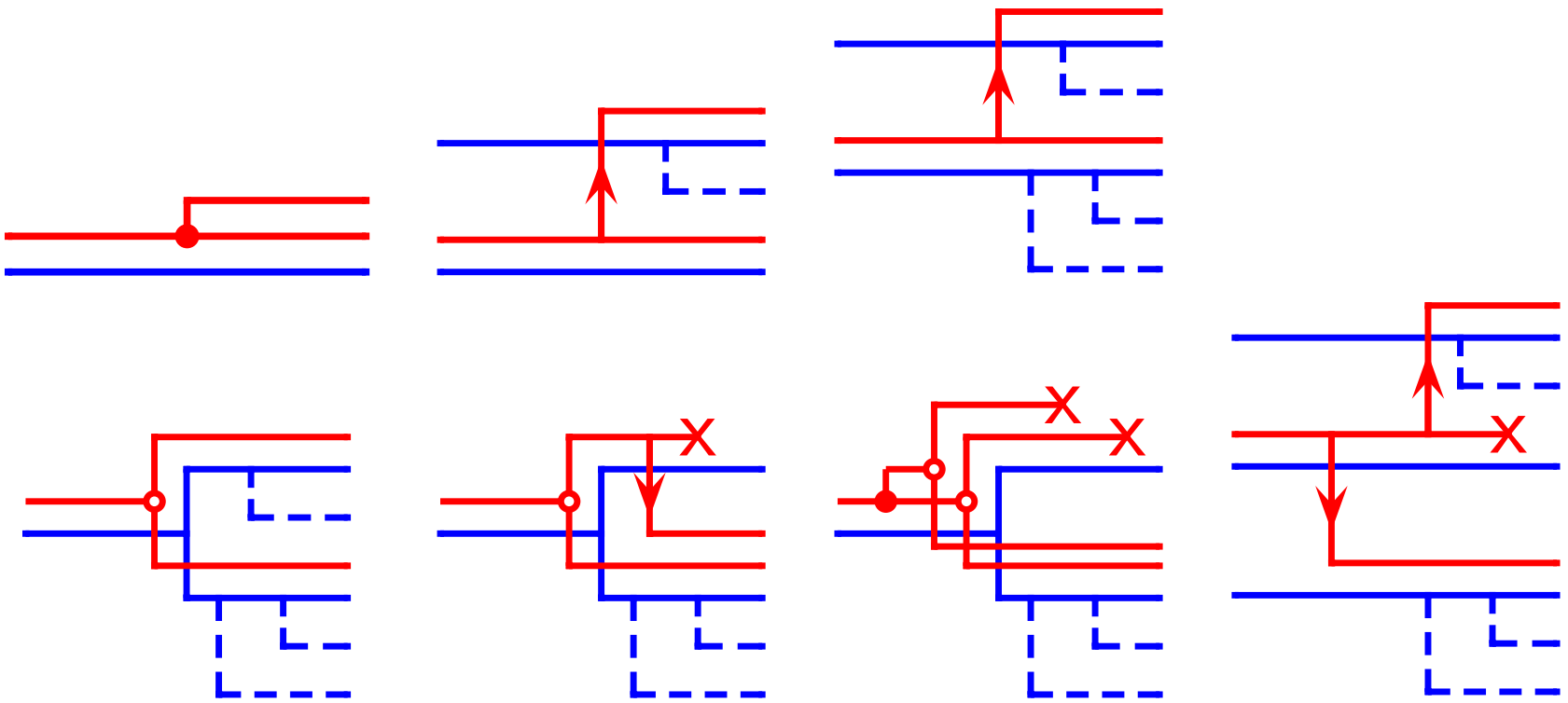
- What is the relationship of h_0 to h_1 and h_2 in the host tree?
- Either self, child/parent, or sister/cousin

Computing $P\left(S \mid H, \mathcal{R}, \theta\right)$



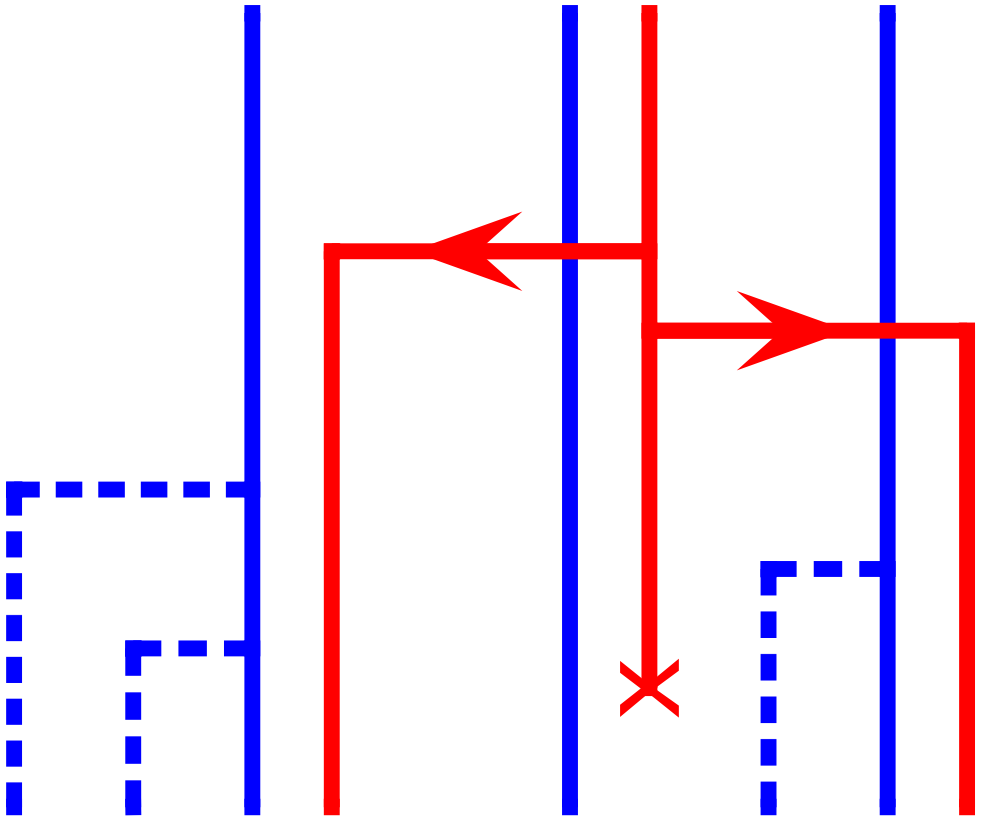
- What is the relationship of h_0 to h_1 and h_2 in the host tree?
- Either self, child/parent, or sister/cousin
- Permutation of two relationships yields potential scenarios

Computing $P\left(S \mid H, \mathcal{R}, \theta\right)$: The Scenarios



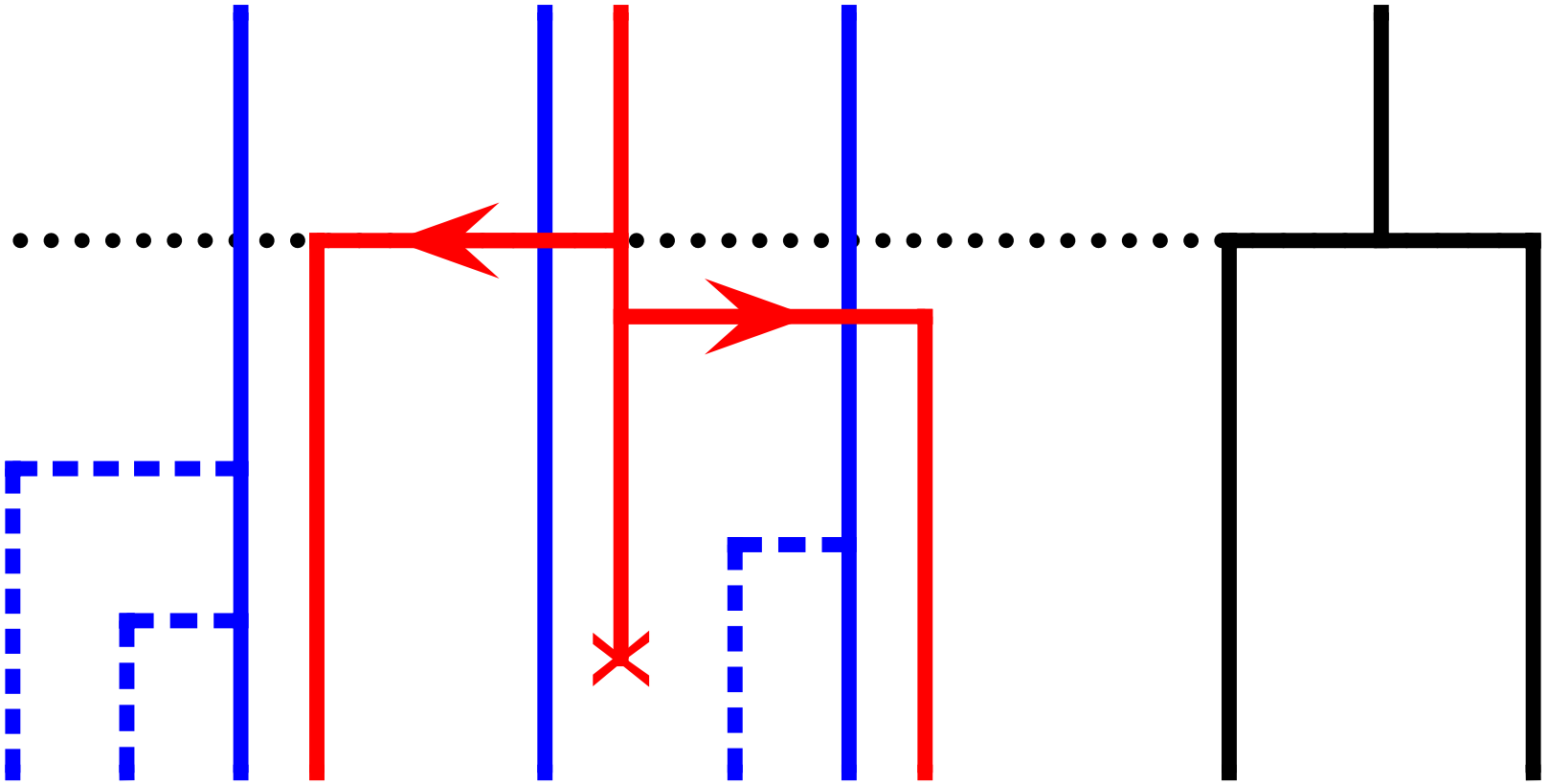
Host / **Symbiont**

Computing $P\left(S \mid H, \mathcal{R}, \theta\right)$: An Example Scenario



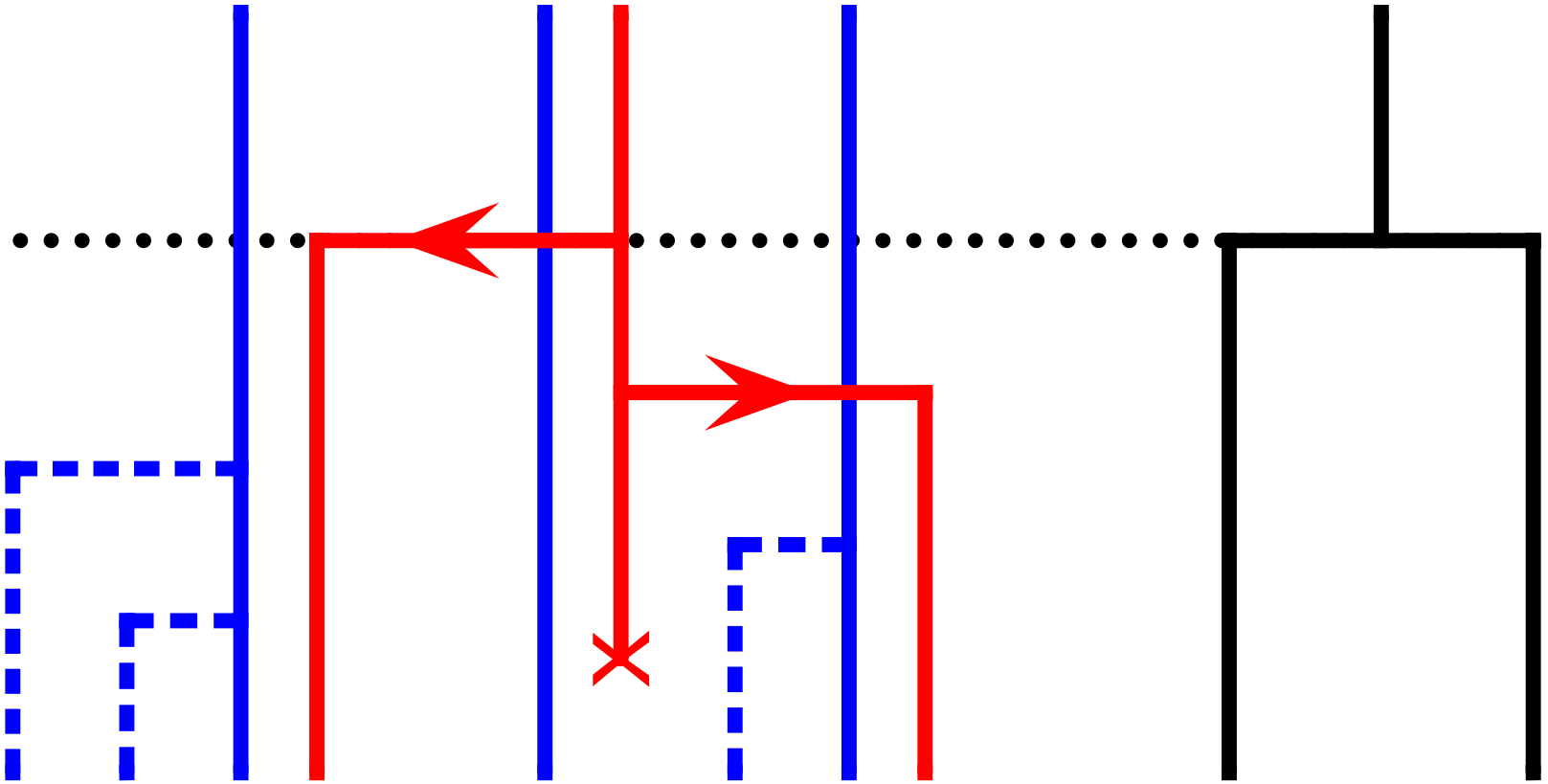
Host / **Symbiont**

Computing $P\left(S \mid H, \mathcal{R}, \theta\right)$: An Example Scenario



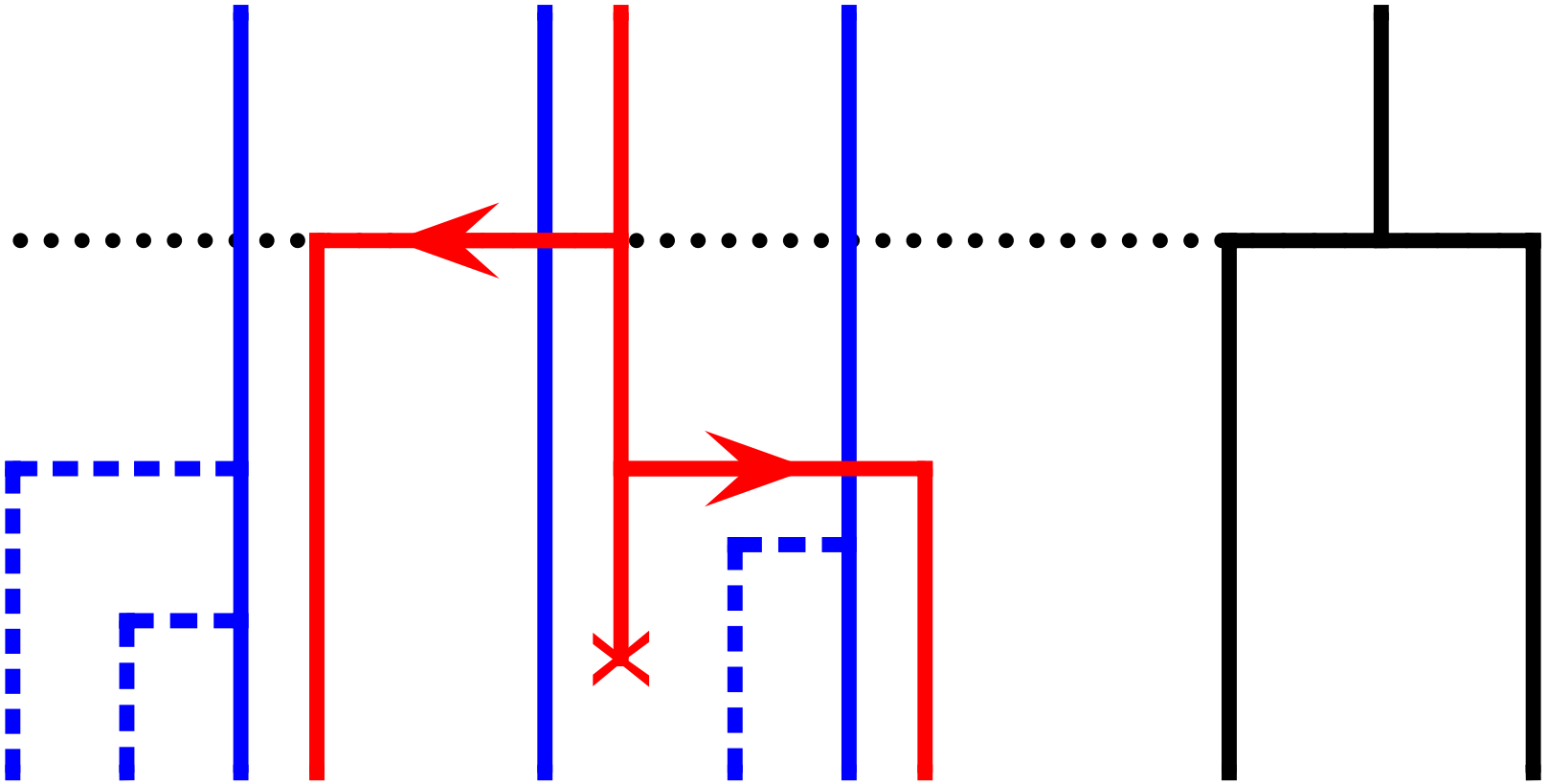
Host / **Symbiont**

Computing $P\left(S \mid H, \mathcal{R}, \theta\right)$: An Example Scenario



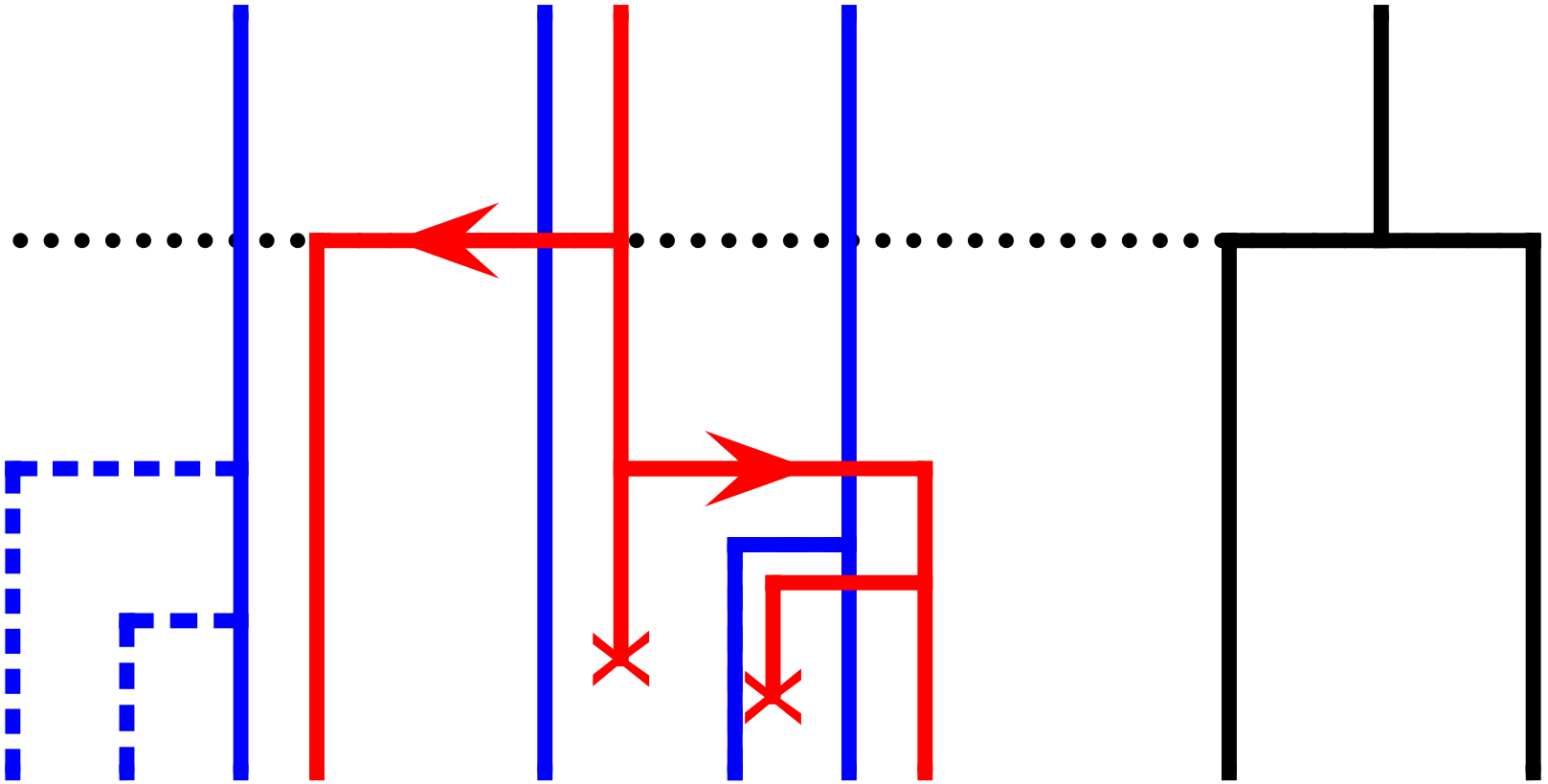
Host / **Symbiont**

Computing $P\left(S \mid H, \mathcal{R}, \theta\right)$: An Example Scenario



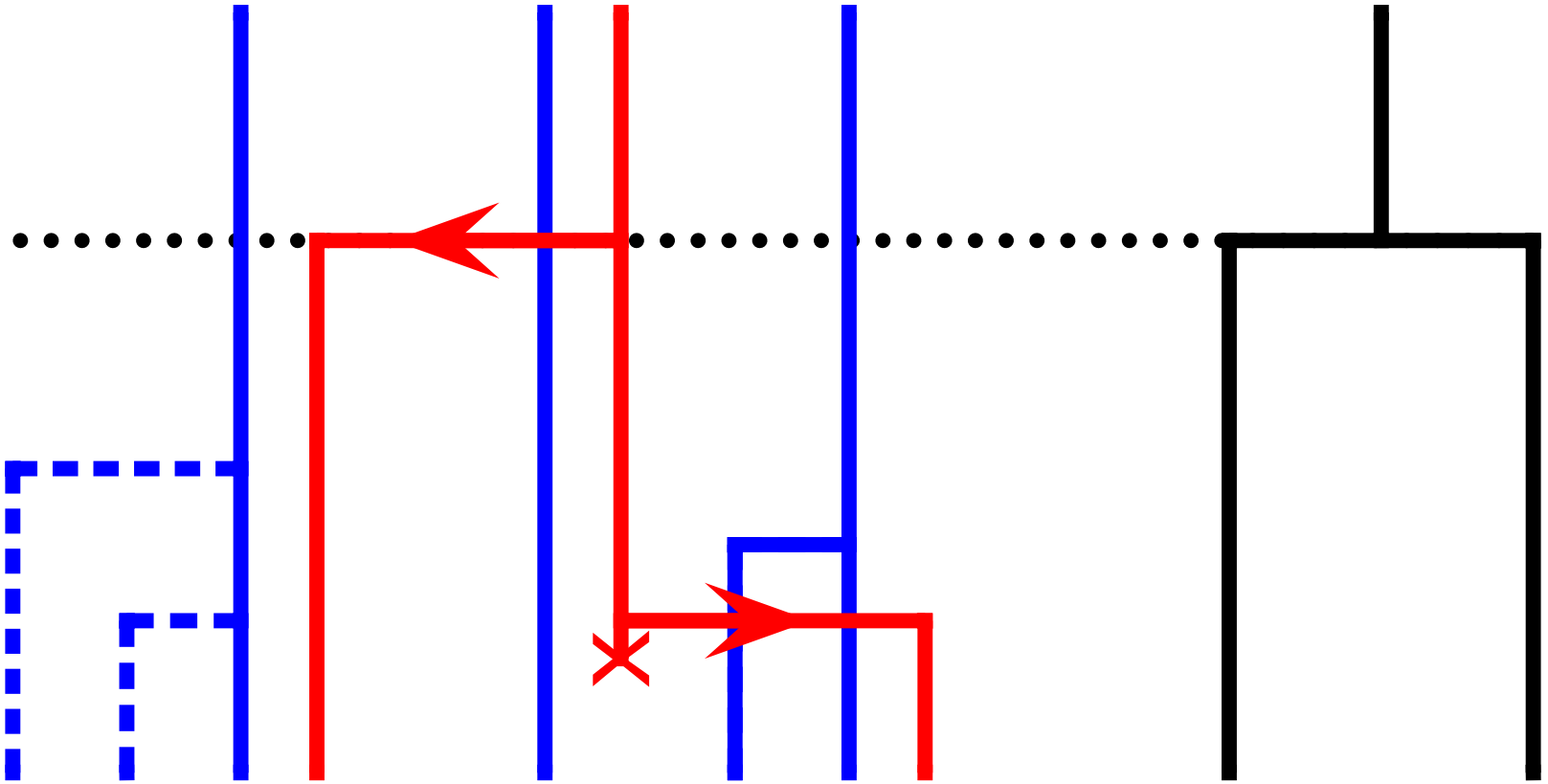
Host / **Symbiont**

Computing $P\left(S \mid H, \mathcal{R}, \theta\right)$: An Example Scenario



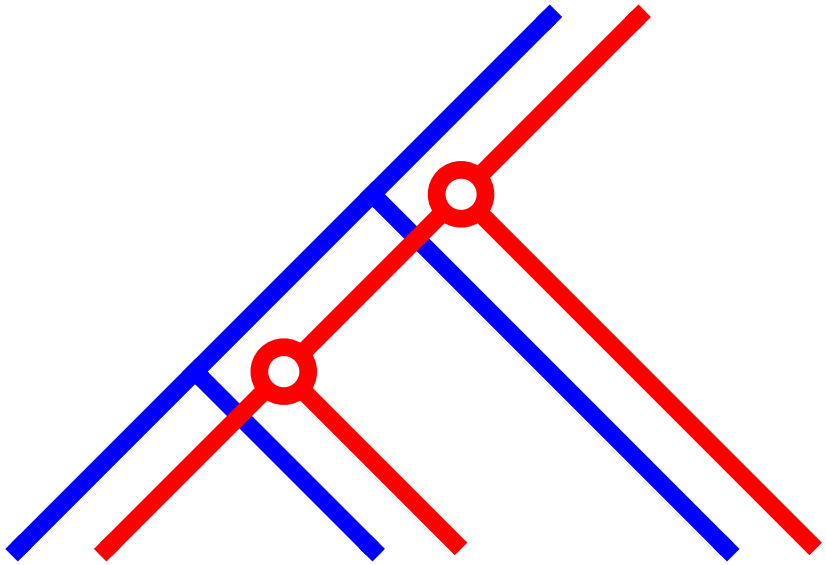
Host / **Symbiont**

Computing $P\left(S \mid H, \mathcal{R}, \theta\right)$: An Example Scenario

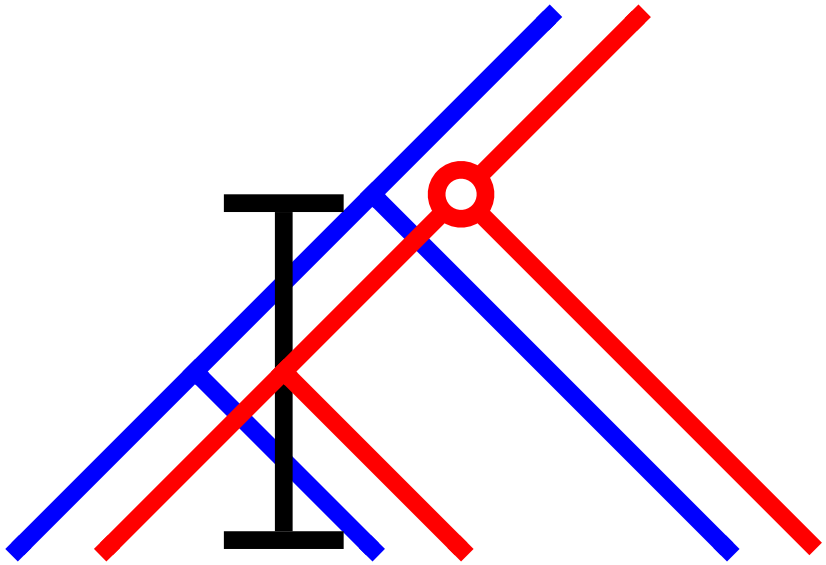


Host / **Symbiont**

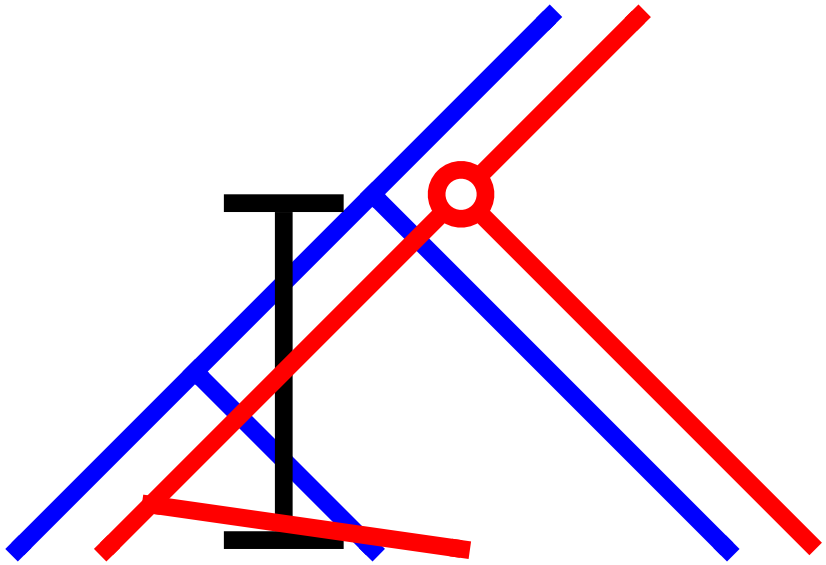
Operating on the Reconciliation



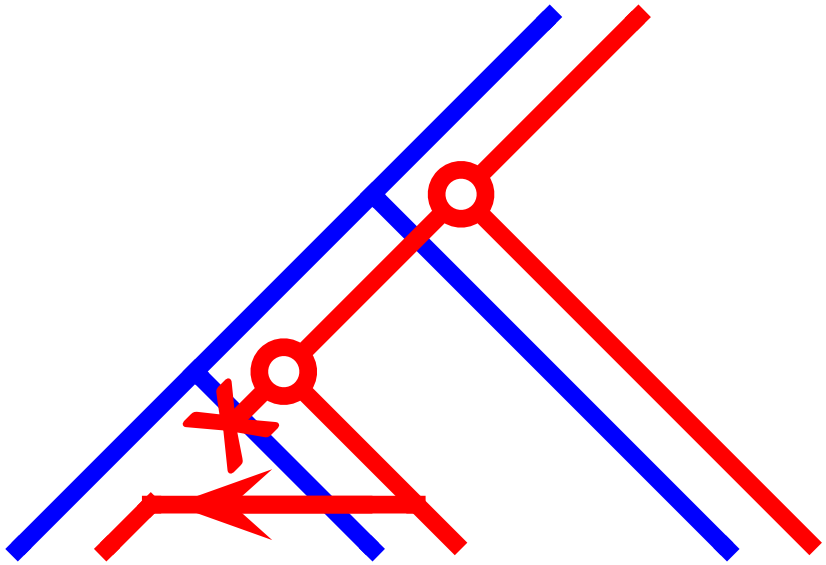
Operating on the Reconciliation



Operating on the Reconciliation



Operating on the Reconciliation



Implementation Details

- $\theta = (\lambda, \tau, \mu)$, rates for duplication, host-switch, and loss

Implementation Details

- $\theta = (\lambda, \tau, \mu)$, rates for duplication, host-switch, and loss
- Introduce rate factor κ represented by clock model
  - Fix $\mu = 1$

Implementation Details

- $\theta = (\lambda, \tau, \mu)$, rates for duplication, host-switch, and loss
- Introduce rate factor κ represented by clock model
  - Fix $\mu = 1$
- Place uniform prior on κ and gamma priors on λ and τ; scale operator on all

Implementation Details

- $\theta = (\lambda, \tau, \mu)$, rates for duplication, host-switch, and loss
- Introduce rate factor κ represented by clock model
  - Fix $\mu = 1$
- Place uniform prior on κ and gamma priors on λ and τ; scale operator on all
- Uniform prior on $\mathcal{R}$, but restricted by leaf–leaf mapping

Implementation Details

- $\theta = (\lambda, \tau, \mu)$, rates for duplication, host-switch, and loss
- Introduce rate factor κ represented by clock model
  - Fix $\mu = 1$
- Place uniform prior on κ and gamma priors on λ and τ; scale operator on all
- Uniform prior on $\mathcal{R}$, but restricted by leaf–leaf mapping
- Implemented as plugin for BEAST1

Simulation Methodology

- Simple simulation pipeline
  1. Host tree generated under constant size coalescent
  2. DNA simulated on host tree under JC69 model
  3. Symbiont tree generated on host tree under described model (three Poisson processes)
  4. DNA simulated on symbiont tree under JC69 for all extant taxa

Simulation Methodology

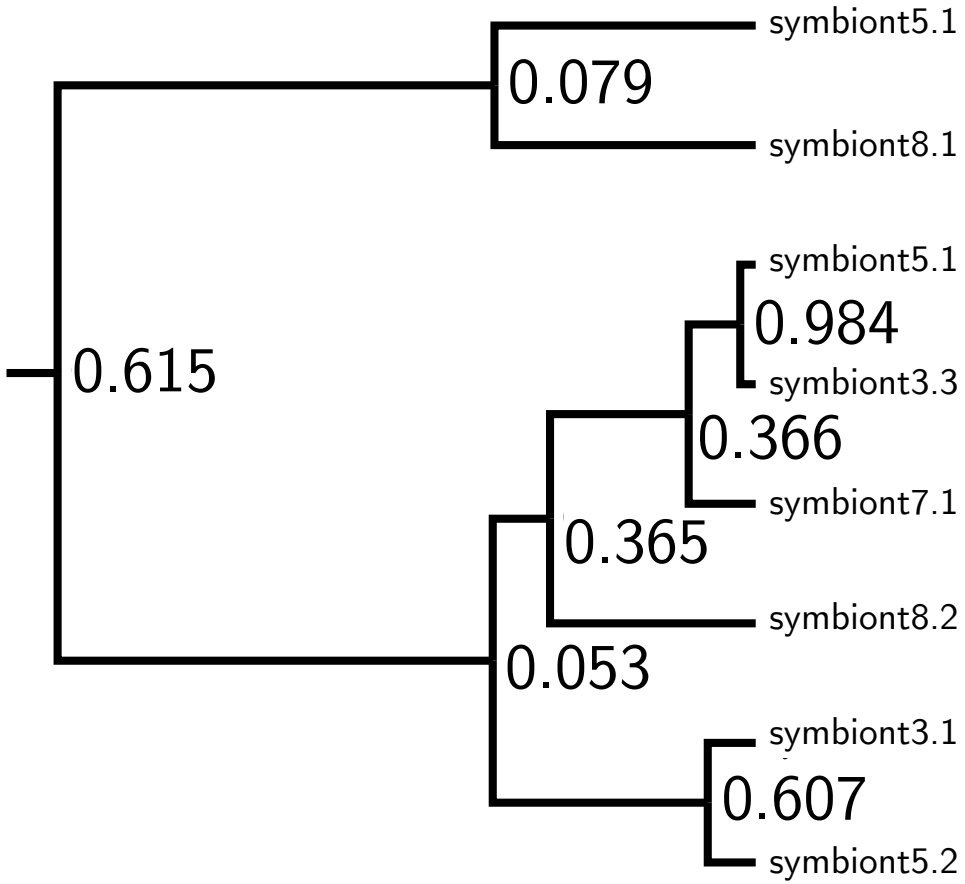
- Simple simulation pipeline
  1. Host tree generated under constant size coalescent
  2. DNA simulated on host tree under JC69 model
  3. Symbiont tree generated on host tree under described model (three Poisson processes)
  4. DNA simulated on symbiont tree under JC69 for all extant taxa
- 1st simulation: 8 host taxa, all event rates 0 (identical trees)
  - Accurate reconstruction with posterior $P \geq 0.99$

Simulation Methodology

- Simple simulation pipeline
  1. Host tree generated under constant size coalescent
  2. DNA simulated on host tree under JC69 model
  3. Symbiont tree generated on host tree under described model (three Poisson processes)
  4. DNA simulated on symbiont tree under JC69 for all extant taxa
- 1st simulation: 8 host taxa, all event rates 0 (identical trees)
  - Accurate reconstruction with posterior $P \geq 0.99$
- 2nd simulation: 8 host taxa, all event rates 1.0, 8 symbionts
  - Trees reconstructed accurately, reconciliation questionable...

Results from Second Simulation

	median	**95% HPD**
λ	1.42	$[0.10, 4.63]$
τ	2.48	$[0.20, 9.18]$
μ	1.60	$[0.11, 7.24]$



Contributions

- Formulated an expression for the posterior probability of a cophylogenetic reconstruction

Contributions

- Formulated an expression for the posterior probability of a cophylogenetic reconstruction
- Developed an algorithm to approximate the probability of a symbiont tree for a reconstruction

Contributions

- Formulated an expression for the posterior probability of a cophylogenetic reconstruction
- Developed an algorithm to approximate the probability of a symbiont tree for a reconstruction
- Implemented the algorithm in an MCMC framework

Open Questions and Problems

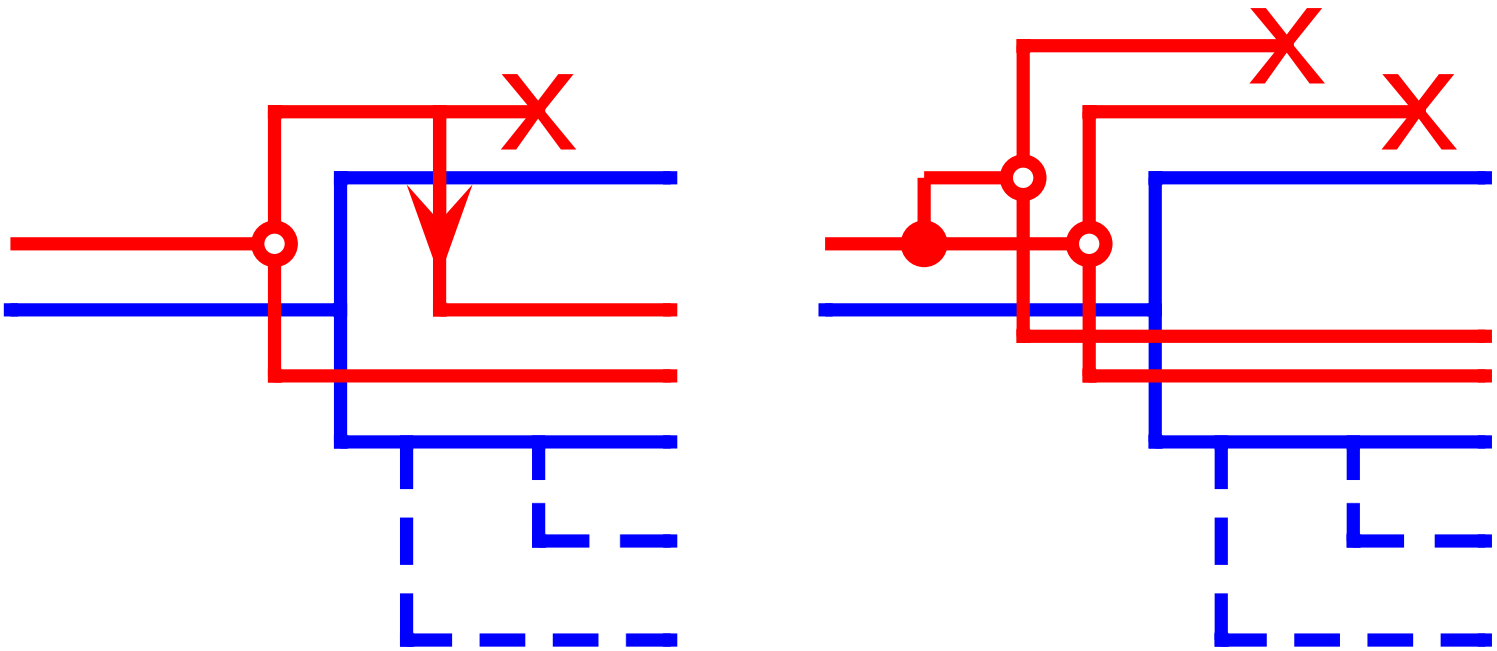
Detecting Cospeciation

- Technically only occurs when host and symbiont speciate at the same time

Open Questions and Problems

Detecting Cospeciation

- Technically only occurs when host and symbiont speciate at the same time
- Important for identifying between some scenarios



Open Questions and Problems

- Is this an appropriate approximation technique?
  - Rigorous definition for an observable event?

Open Questions and Problems

- Is this an appropriate approximation technique?
  - Rigorous definition for an observable event?
- Can we place an informative prior on the reconciliation?
  - $P(H, \mathcal{R}, \theta) = P(\mathcal{R} \mid H, \theta) P(H) P(\theta)$
  - We have a node–node mapping $\mathcal{R}$, but do not know nodal relationships for one tree

Open Questions and Problems

- Is this an appropriate approximation technique?
  - Rigorous definition for an observable event?
- Can we place an informative prior on the reconciliation?
  - $P(H, \mathcal{R}, \theta) = P(\mathcal{R} \mid H, \theta) P(H) P(\theta)$
  - We have a node–node mapping $\mathcal{R}$, but do not know nodal relationships for one tree
- What is the effect of the operator on mixing?
  - Cophylogeny model substantially fragments posterior landscape—a wrong move makes the reconciliation invalid

Open Questions and Problems

- Is this an appropriate approximation technique?
  - Rigorous definition for an observable event?
- Can we place an informative prior on the reconciliation?
  - $P(H, \mathcal{R}, \theta) = P(\mathcal{R} \mid H, \theta) P(H) P(\theta)$
  - We have a node–node mapping $\mathcal{R}$, but do not know nodal relationships for one tree
- What is the effect of the operator on mixing?
  - Cophylogeny model substantially fragments posterior landscape—a wrong move makes the reconciliation invalid
- How best can we evaluate the model performance via simulation studies?

Open Questions and Problems

- Are the event rates better predicted by the symbiont or host?

Open Questions and Problems

- Are the event rates better predicted by the symbiont or host?
- Model for host speciating independently of symbiont (a.k.a. failure to diverge event)?

Open Questions and Problems

- Are the event rates better predicted by the symbiont or host?
- Model for host speciating independently of symbiont (a.k.a. failure to diverge event)?
- Can we consider preferential host-switching in the model?

Open Questions and Problems

- Are the event rates better predicted by the symbiont or host?
- Model for host speciating independently of symbiont (a.k.a. failure to diverge event)?
- Can we consider preferential host-switching in the model?
- Can we consider geography in the model?
  - An associated host and symbiont must be cohabiting

Open Questions and Problems

- Are the event rates better predicted by the symbiont or host?
- Model for host speciating independently of symbiont (a.k.a. failure to diverge event)?
- Can we consider preferential host-switching in the model?
- Can we consider geography in the model?
  - An associated host and symbiont must be cohabiting
- How can we visualise the reconstruction?
  - Reconciling the trees does not recover the events
  - Several uncertainties in event timing, etc.

Open Questions and Problems

- Are the event rates better predicted by the symbiont or host?
- Model for host speciating independently of symbiont (a.k.a. failure to diverge event)?
- Can we consider preferential host-switching in the model?
- Can we consider geography in the model?
  - An associated host and symbiont must be cohabiting
- How can we visualise the reconstruction?
  - Reconciling the trees does not recover the events
  - Several uncertainties in event timing, etc.
- Can we test coevolutionary theories, e.g. GMTC or escape-and-radiate?

Acknowledgements

- My mentors, Dr. Yi-Chieh Jessica Wu, Rachel Sealfon, and Prof. Mukul Bansal
- Andrew Brownjohn, Jon Sanders, Prof. Ran Libeskind-Hadas, Hayden Metsky, and Prof. Manolis Kellis, for helpful discussions
- Dr. Susan Offner, for inspiring me