# Algorithmic issues in (co)phylogenetic analysis

blerina sinaimeri

# Motivation

**Different systems “coevolve”**

- hosts and their parasites or pathogens
- whole organisms and their genes
- geographical areas and the species which inhabit them.
- cultural traditions and populations

Same model

**Host-Parasite associations**

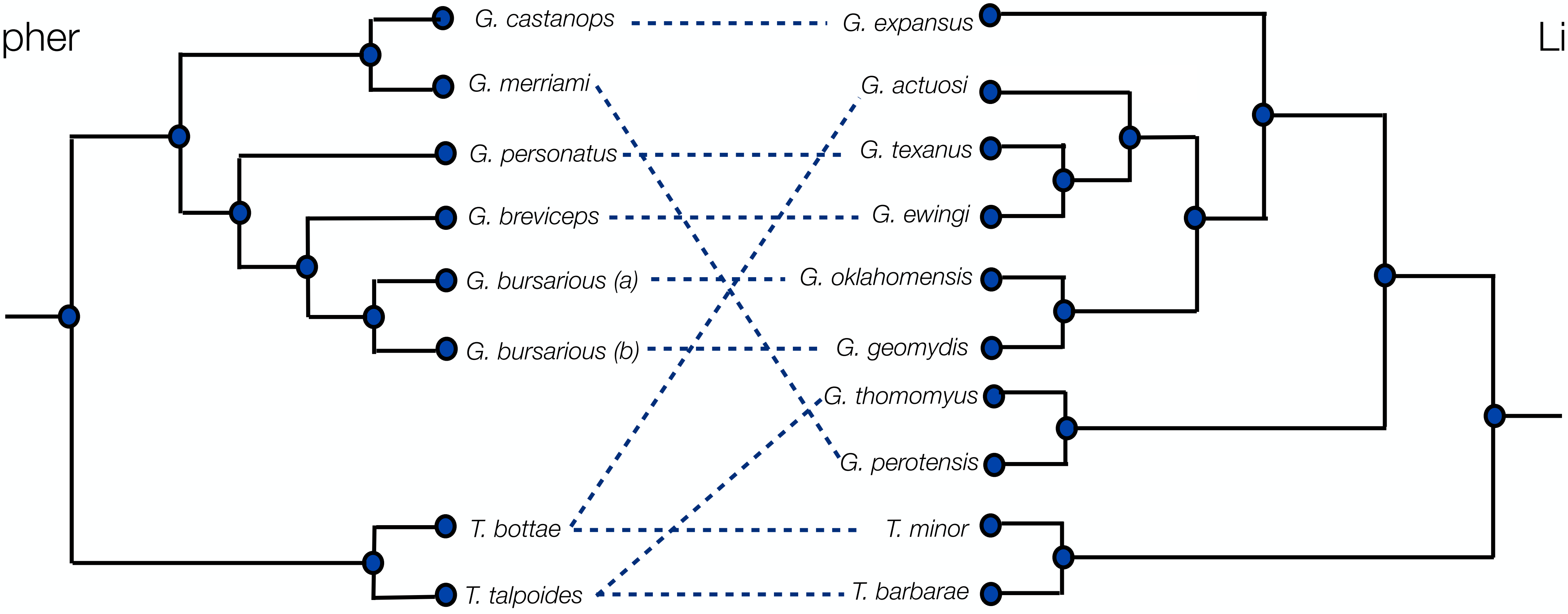
- about 75% of emergent human diseases are zoonoses, that  is, they switched hosts from other species into humans
- determine the rates of evolution in hosts and parasite
- determine how long is the association between host and parasite

# The instance of the problem

**Co-evolution**

Gopher

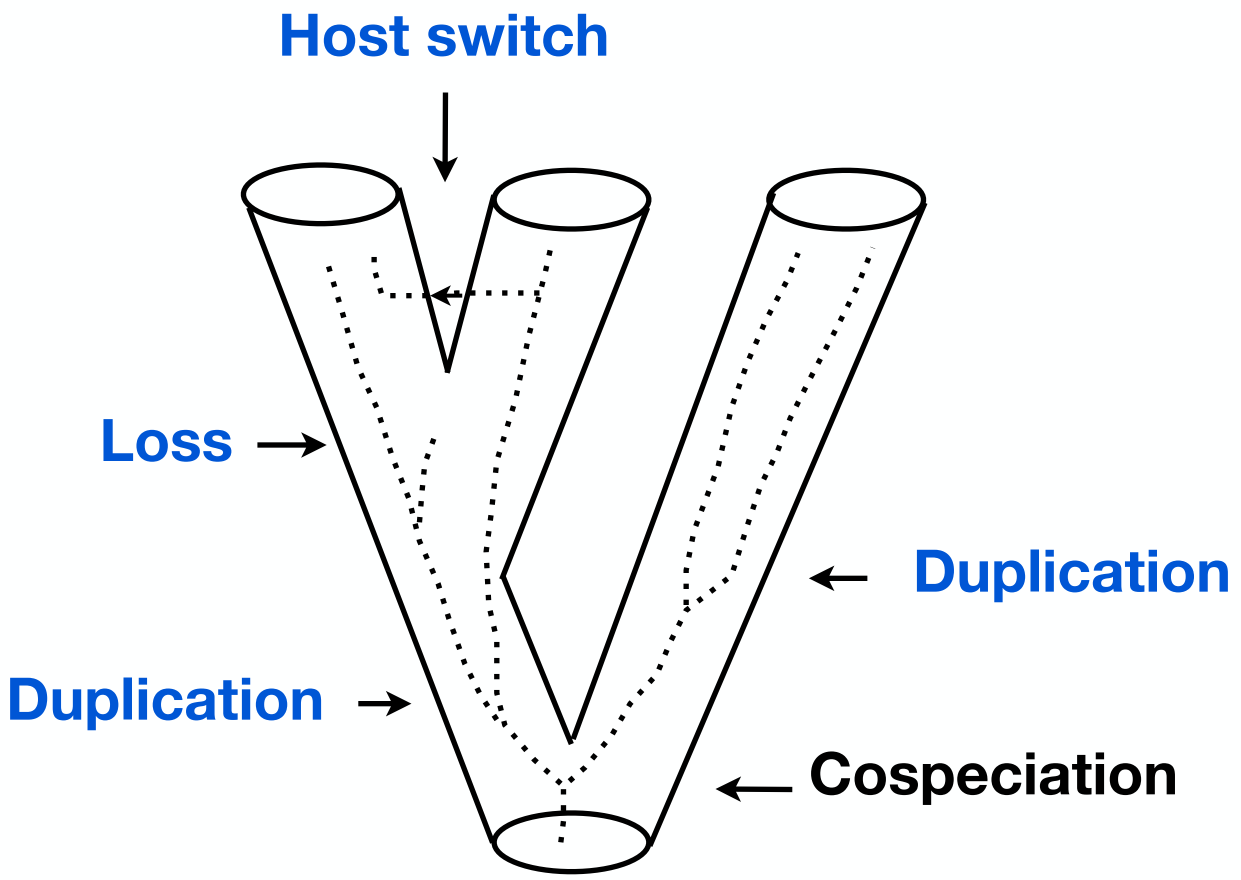
Lice



*adapted from Penn, Dustin J(Apr 2001) Coevolution: Host–Parasite. In: eLS. John Wiley & Sons Ltd, Chichester. http://www.els.net

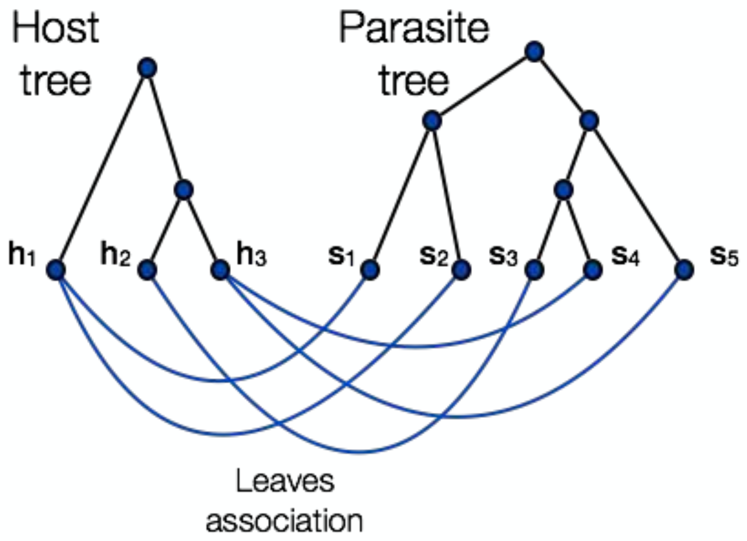
# Reconciliation method

**Co-phylogeny reconstruction problem**



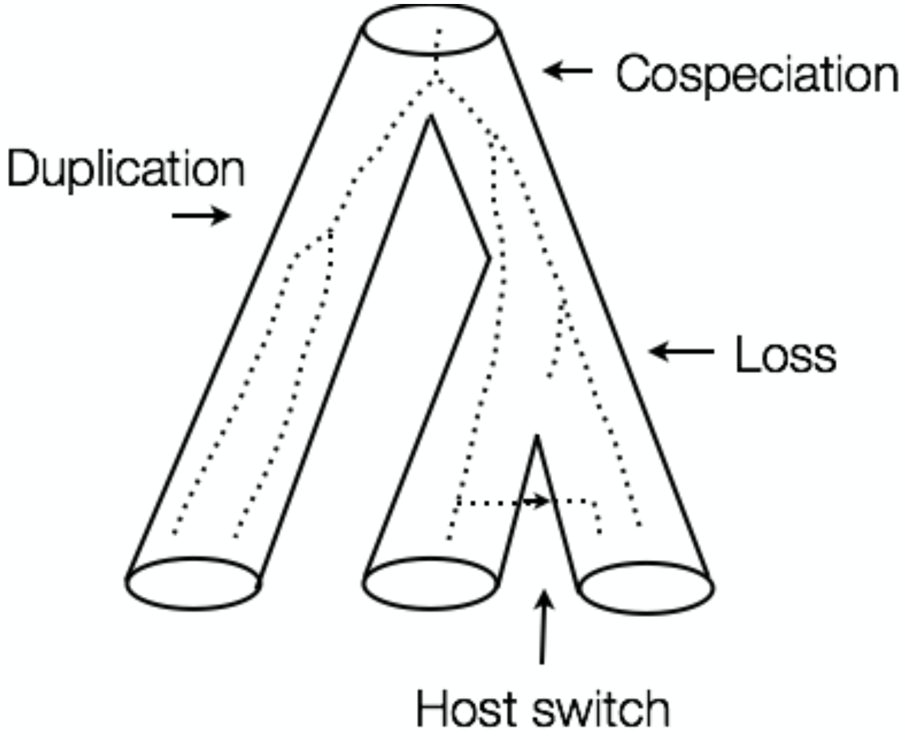
# State-of-art: reconciliation method

## The Input



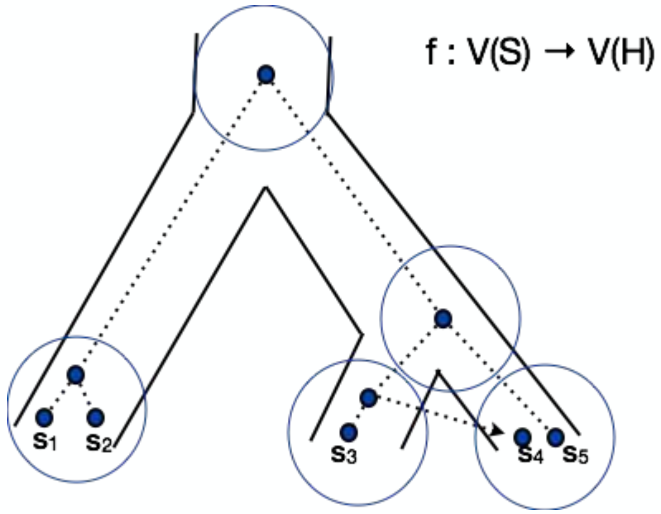
**Two trees and the association function between their leaves.**

## What we're looking for



**The four types of possible events.**

## The mapping function



Selecting the best solution: assign a cost to each of the four types of events and then minimize the total cost.

# Modeling the events

The mapping $f$ induces a partition of $V(P)$ into three sets:

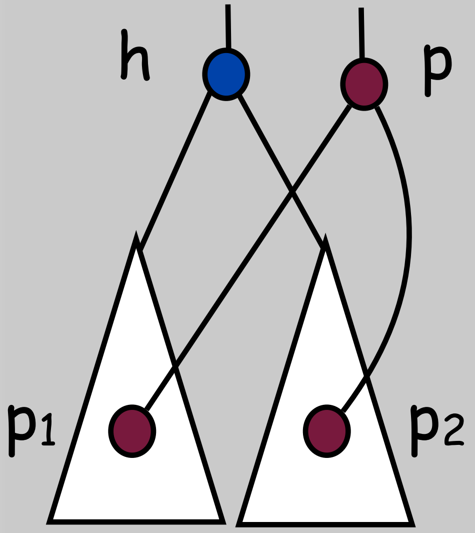
- $\Sigma$ → co-speciations
- $\Delta$ → duplications
- $\Theta$ → host-switches

# Modeling the events

The mapping $f$ induces a partition of $V(P)$ into three sets:

- $\Sigma$ → co-speciations
- $\Delta$ → duplications
- $\Theta$ → host-switches

- Co-speciation



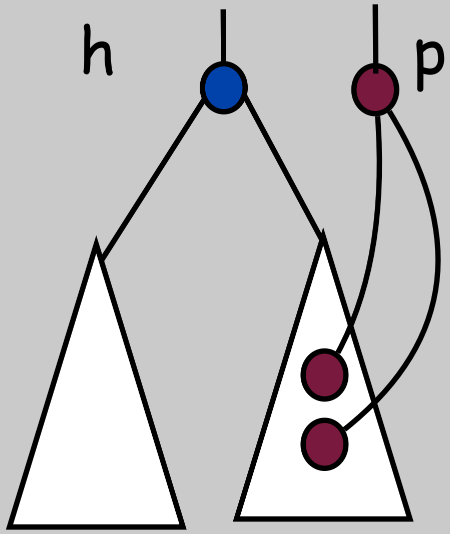
$lca( f(p_1), f(p_2) )= f(p)$ and $f(p_1)$ and $f(p_2)$ are incomparable.

# Modeling the events

The mapping $f$ induces a partition of $V(P)$ into three sets:

- Σ → co-speciations
- Δ → duplications
- Θ → host-switches

- Duplication



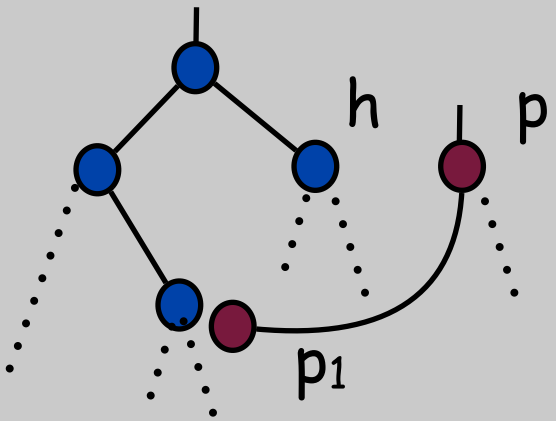
$lca\ (f(p_1), f(p_2)) \in \{f(p_1), f(p_2)\}$

# Modeling the events

The mapping $f$ induces a partition of $V(P)$ into three sets:

- $\Sigma$ → co-speciations
- $\Delta$ → duplications
- $\Theta$ → host-switches

- Host-switch

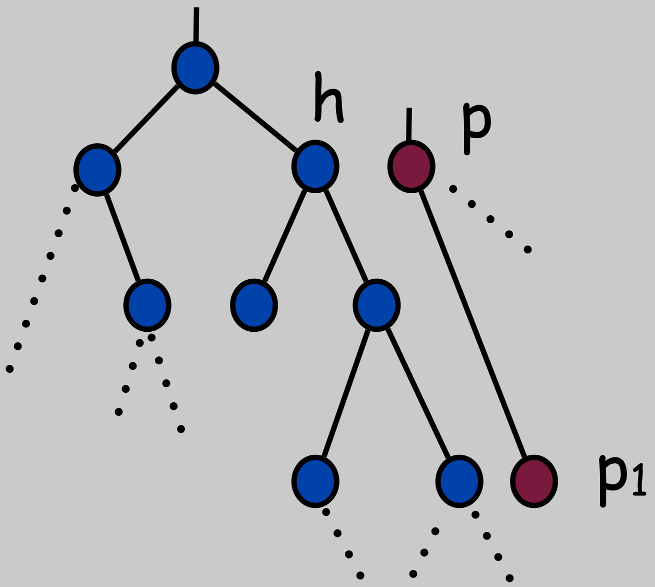


$\mathrm{lca}\ (f(p_1), f(p)) \neq f(p)$

# Modeling the events

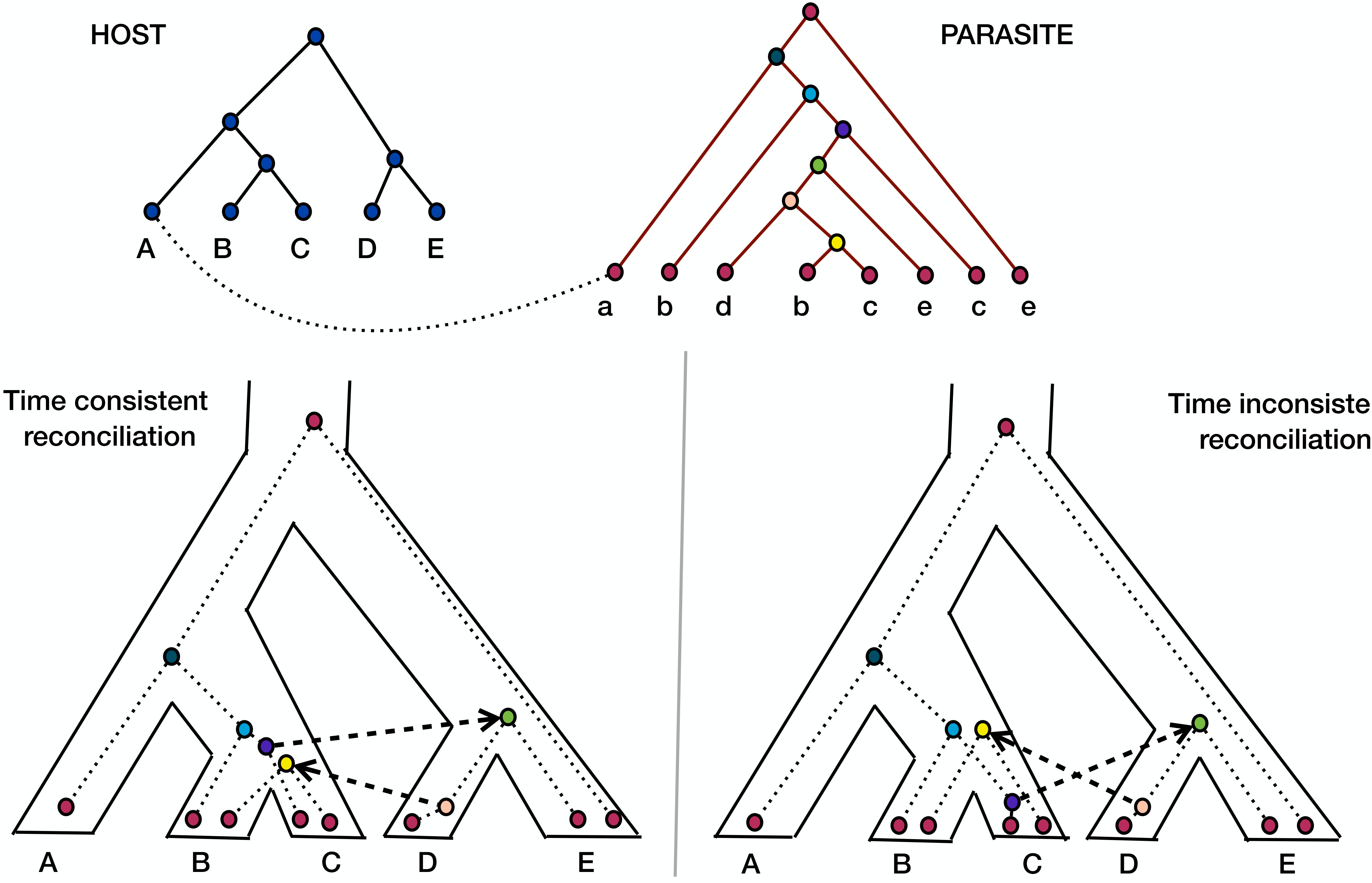
We can define a function $\alpha(f)$ that gives the number losses induced by the mapping $f$.

- Loss



the edge $(p, p_1)$ contributes with $1$ loss.

# Everything is against us.....



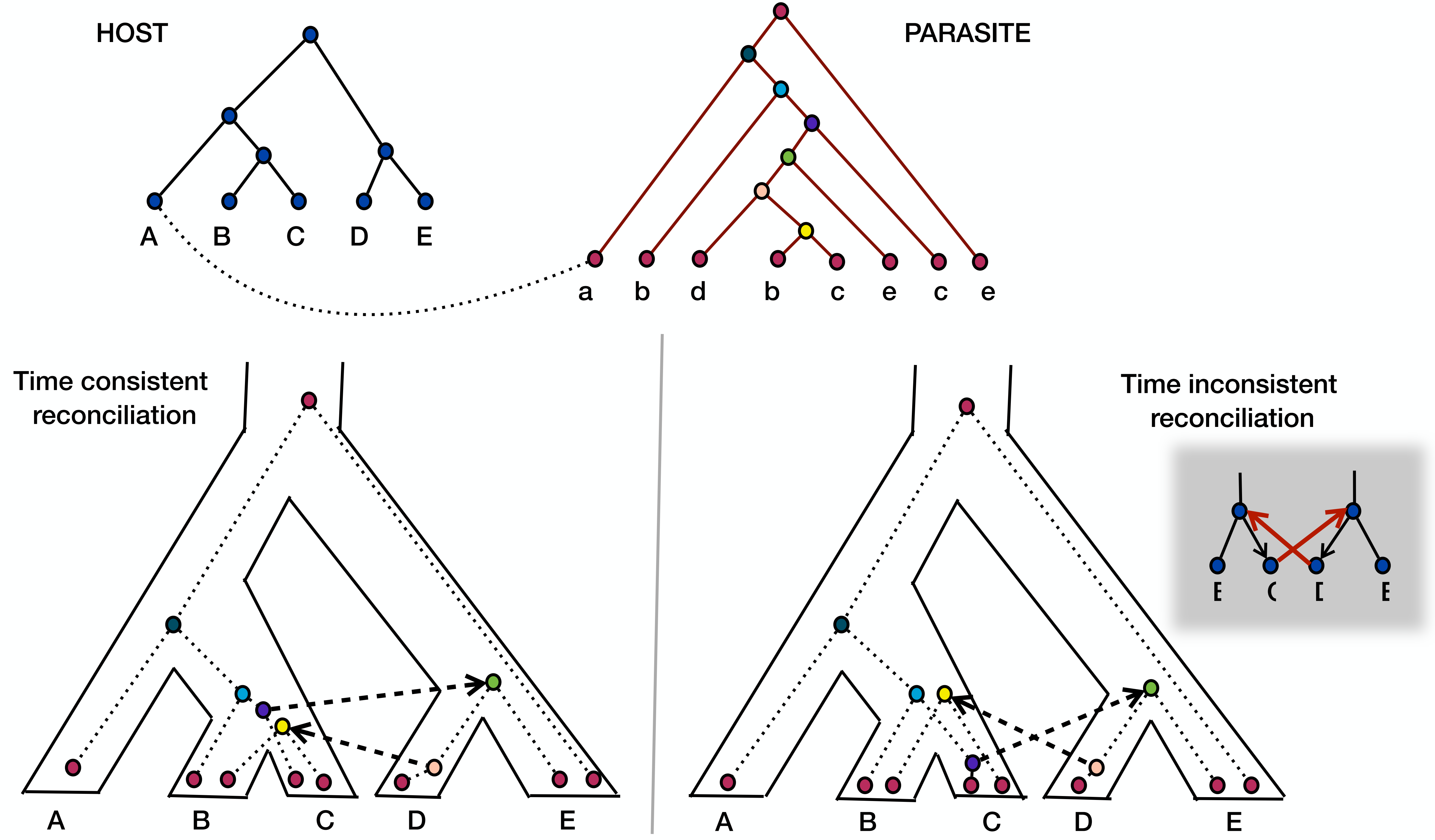
HOST

PARASITE

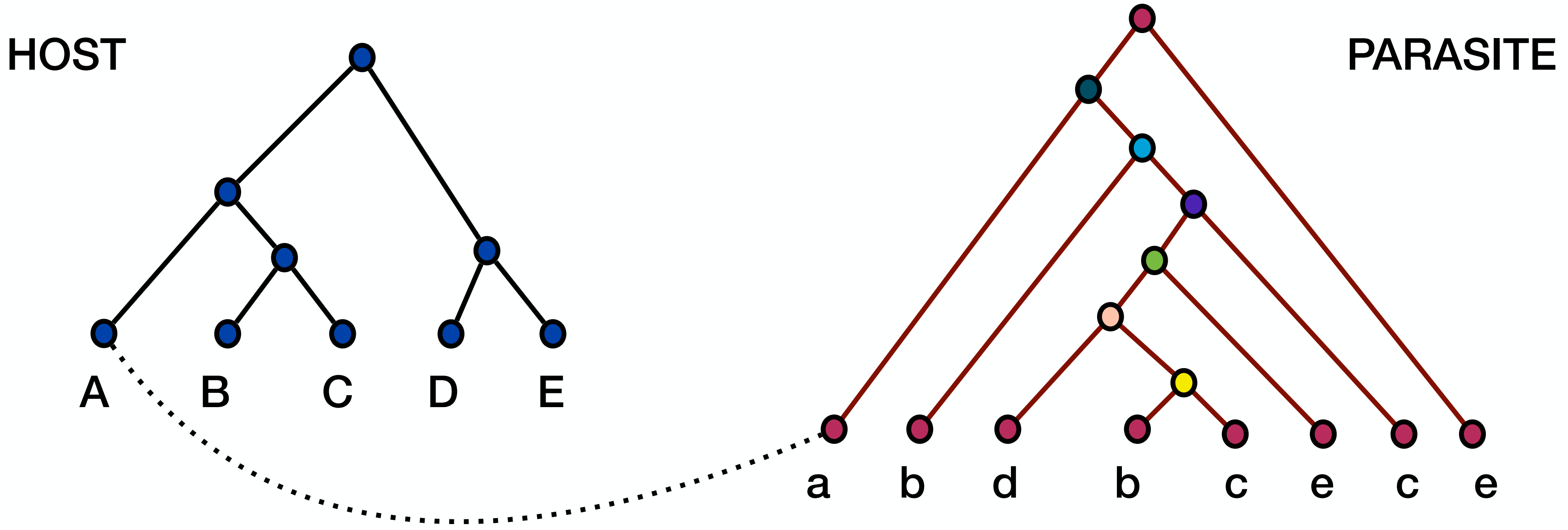
Time consistent reconciliation

Time inconsistent reconciliation

# Everything is against us.....

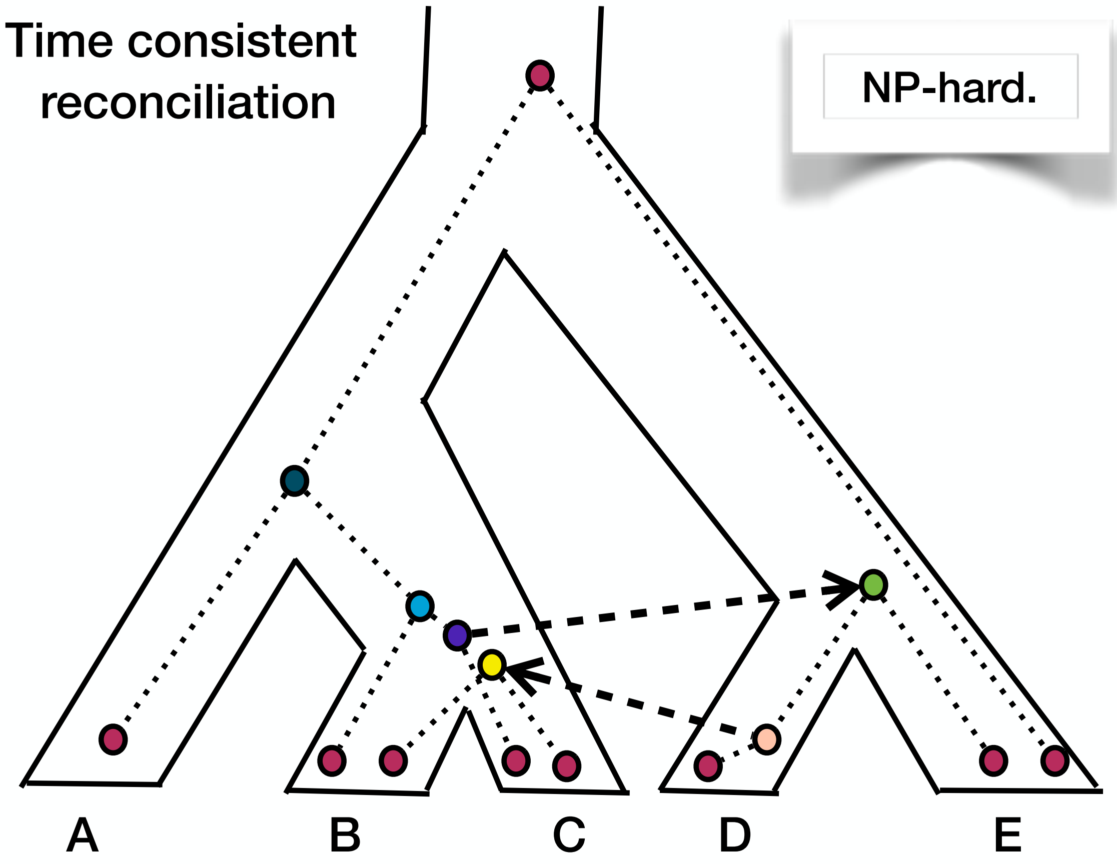

HOST
A
B
C
D
E
PARASITE
a
b
d
b
c
e
c
e
Time consistent reconciliation
A
B
C
D
E
Time inconsistent reconciliation
A
B
C
D
E

# Everything is against us.....



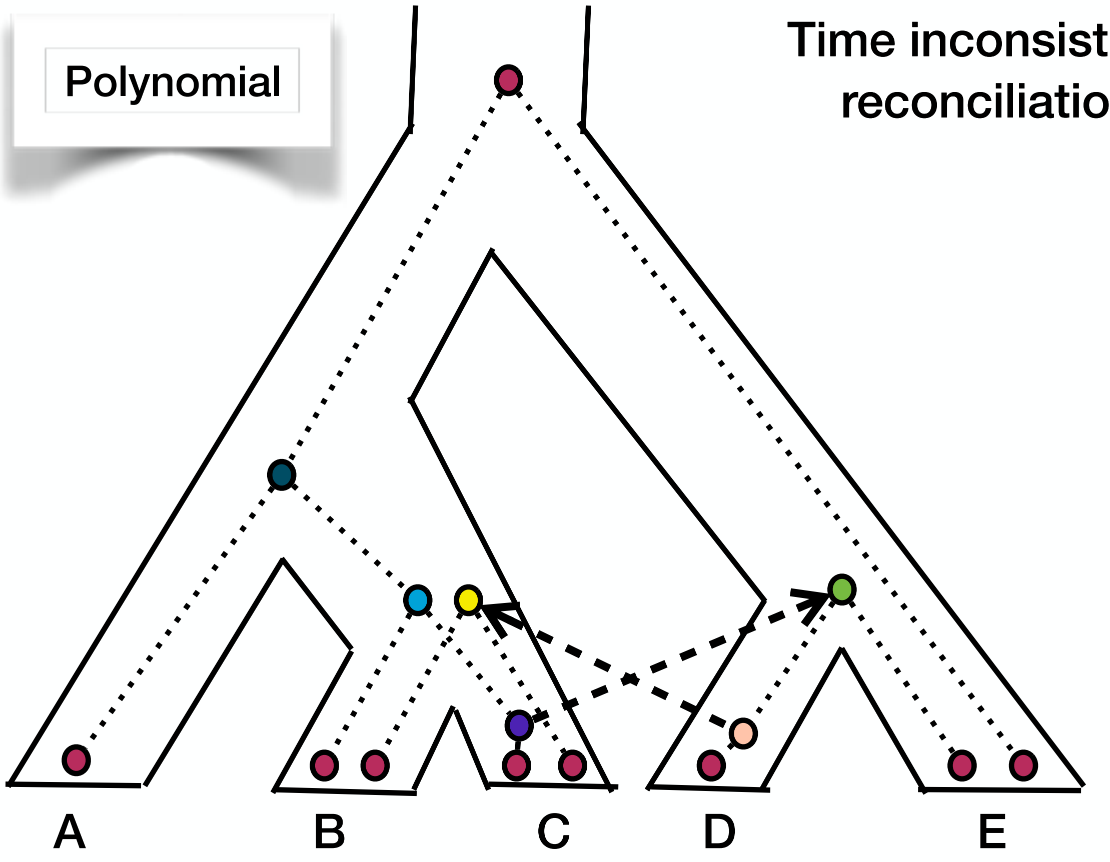
Time consistent reconciliation

NP-hard.



Polynomial

Time inconsistent reconciliation



# Everything is against us.....

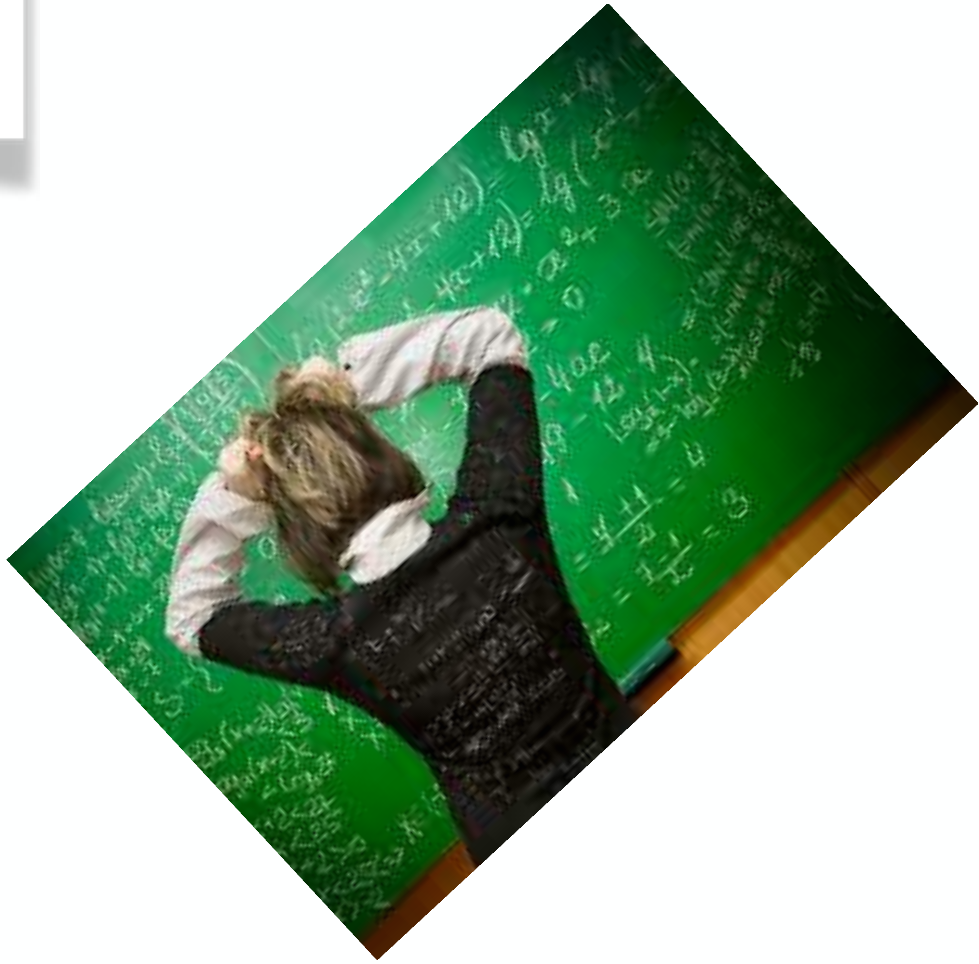
Finding an optimal reconciliation is NP-hard.

The complexity arises from the difficulty of separating possible from impossible host switches combinations.

# Everything is against us.....

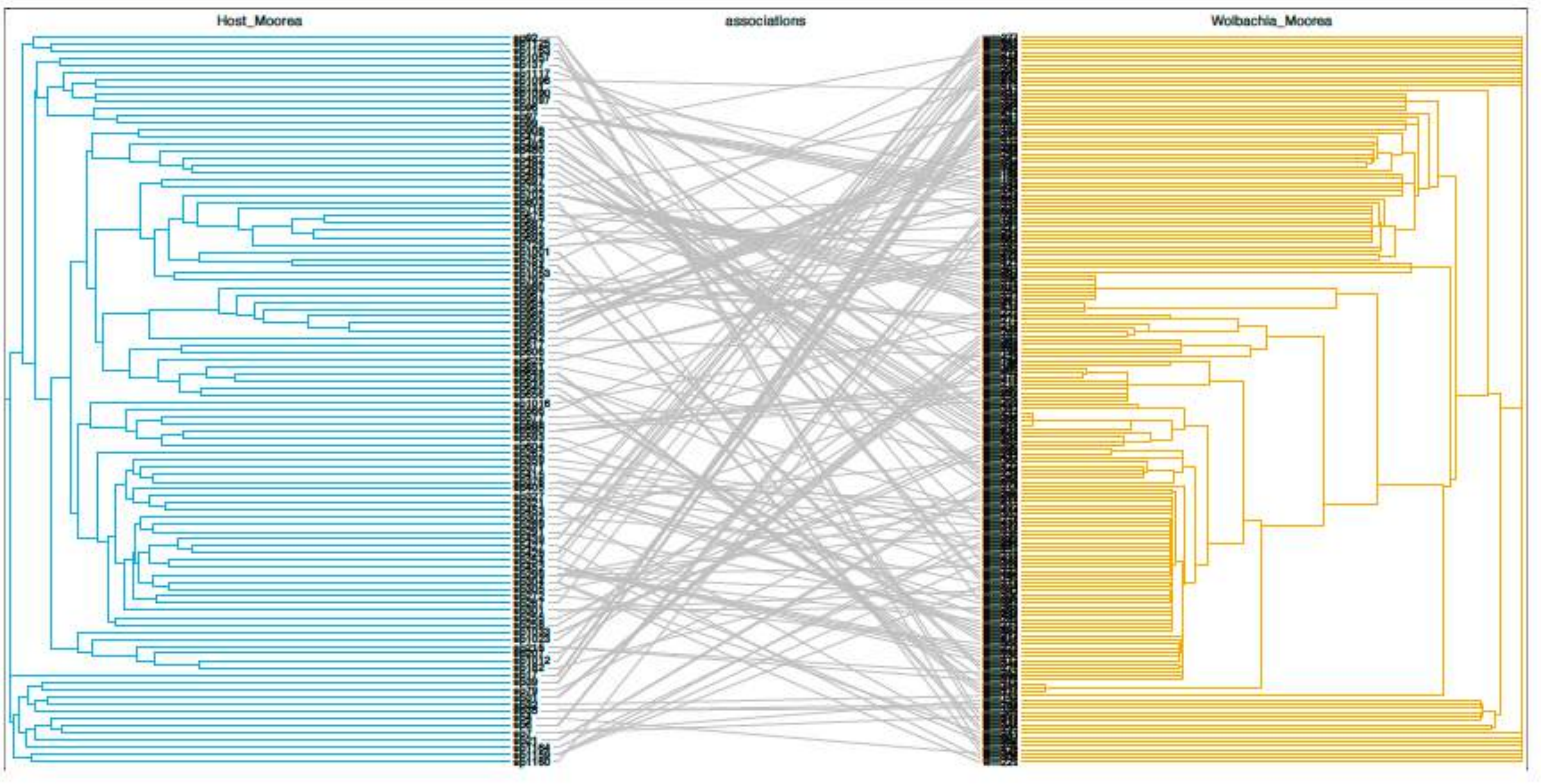
Generate all the optimal reconciliations.

- The number of optimal reconciliations increases rapidly even for small trees.
- The size of the trees can be large.



# Real data

A sample with hundreds of arthropods and the Wolbachia infecting them.
Data from Patricia Simões, collected in Tahiti, Moorea, Raiatea.



Wolbachia in Moorea

# Our contribution so far

A **polynomial delay** algorithm for generating all the optimal reconciliations.

Basic idea:

- Fill a dynamic programming matrix with additional information for the exhaustive traceback.

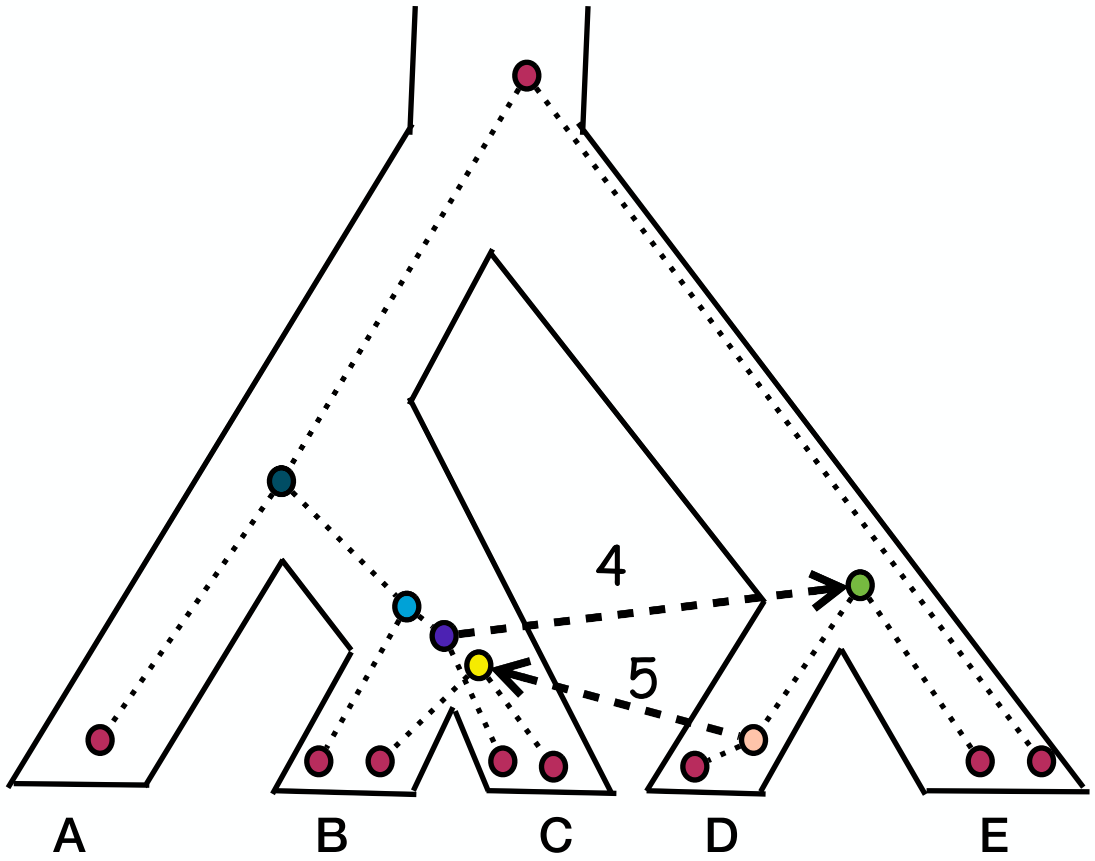
Problems

- No time-feasible solutions
- Too many time-feasible solutions

# Bounded switch problem

**k-switch Problem:**

Given $H$, $P$, $\varphi$, $\underline{c}$, and an integer k find an optimal reconcilation in which all the host switches have a distance bounded by k.



# Bounded switch problem

**k-switch Problem:**

Given $H$, $P$, $\varphi$, $\underline{c}$, and an integer k find an optimal reconcilation in which all the host switches have a distance bounded by k.

- host-switches only between closely related species.
- No time-feasible solutions ⇒ decrease $k$.
- Too many time-feasible solutions ⇒ decrease $k$ maintaining the same optimal cost.

**Open Problem**

What is the complexity of the k-switch problem in the acyclic case?

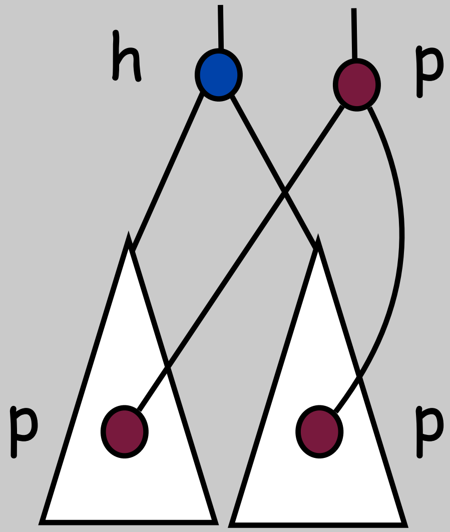
# Exercise 2

- Given two phylogenetic trees is it possible to find a reconciliation without host-switches? Without duplications?

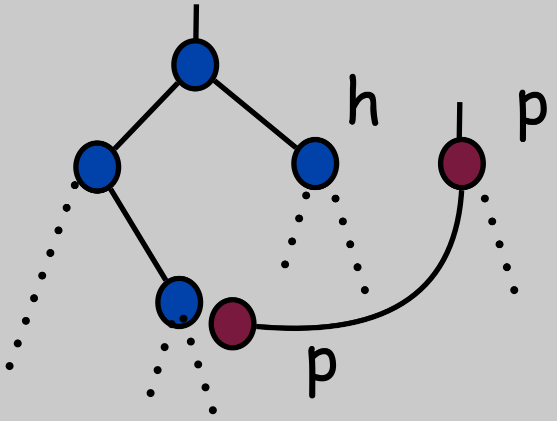
# Current work

## Other types of events

- Failure to diverge



- Spread



# Open Problems

**More realistic models**

- the cost values influence the optimal solution
- multiple hosts - multiple parasites (communities)

# Sequential speciation

Leaf cutter ants

# phylogenetic forests

# Sequential speciation

Leaf cutter ants