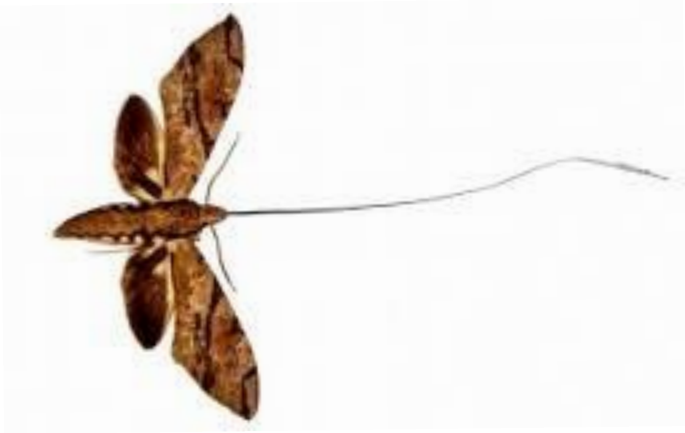


Reconciling phylogenetic trees

blerina sinaimeri

Workshop CWI - INRIA 2018

Interspecific interactions

Plant diversity

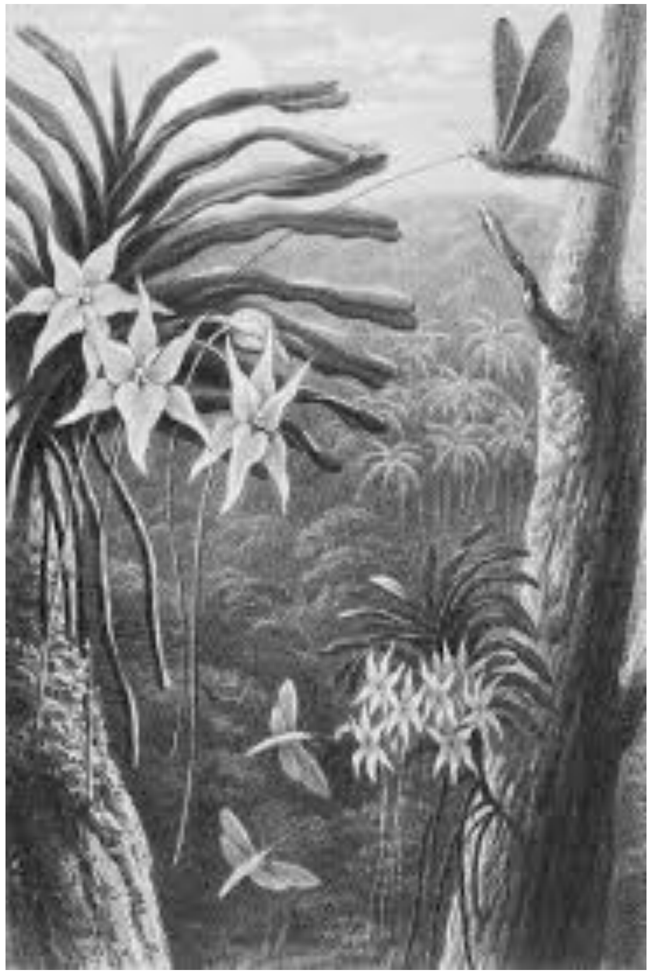
Mimicry

Human Microbiota

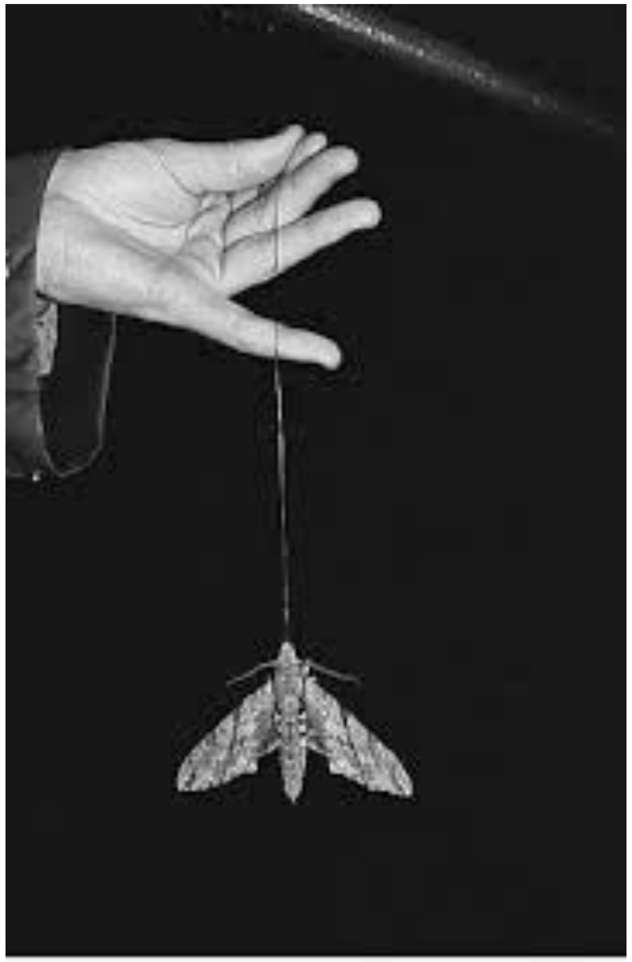
Parasitism

Mutualism

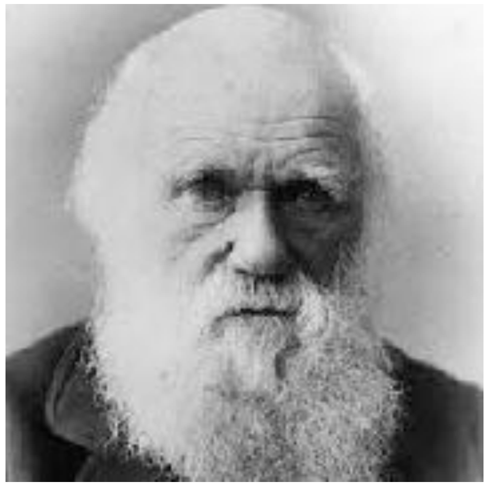
Coevolution



Star orchid (Angraecum sesquipedale) from Madagascar, which has 25 cm long flower spur



The Hawk moth Xanthopan morganii praedicta



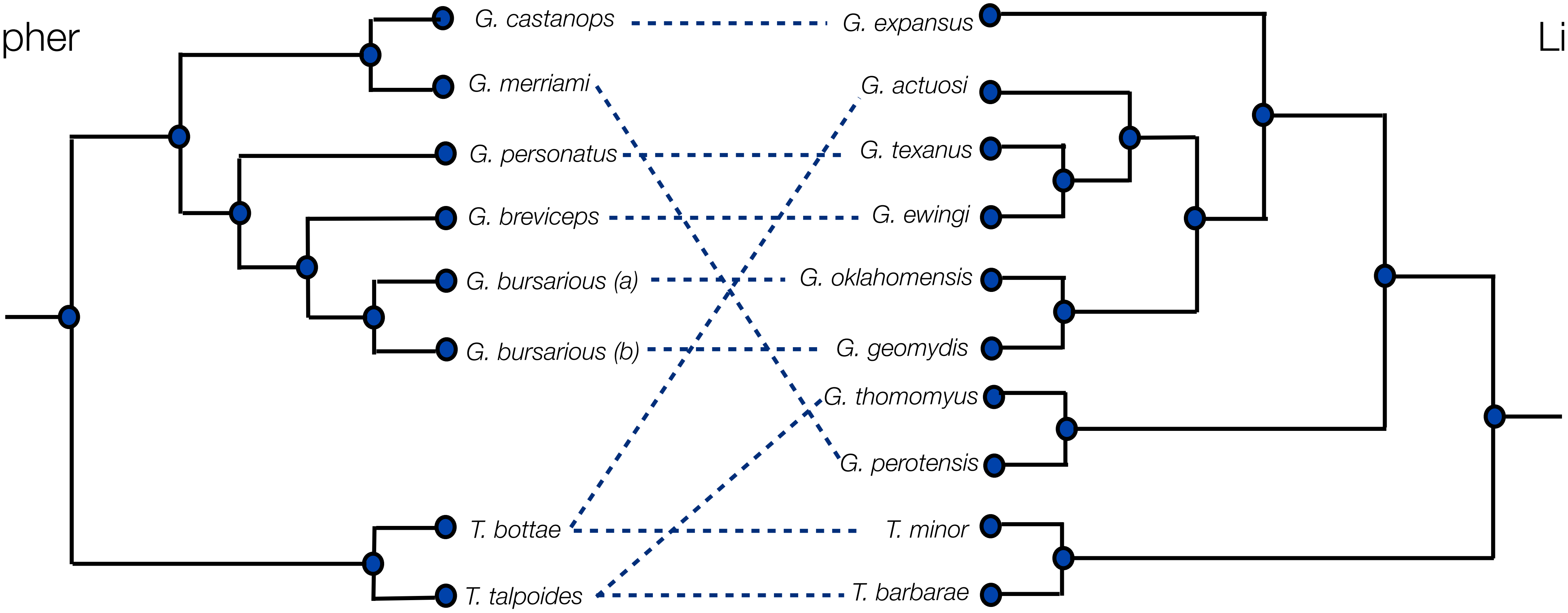
"Thus I can understand how a flower and a bee might slowly become, either simultaneously or one after the other, modified and adapted to each other in the most perfect manner, by the continued preservation of all the individuals which presented slight deviations of structure mutually favourable to each other."
— Charles Darwin, The Origin of Species

The instance of the problem

Co-evolution

Gopher

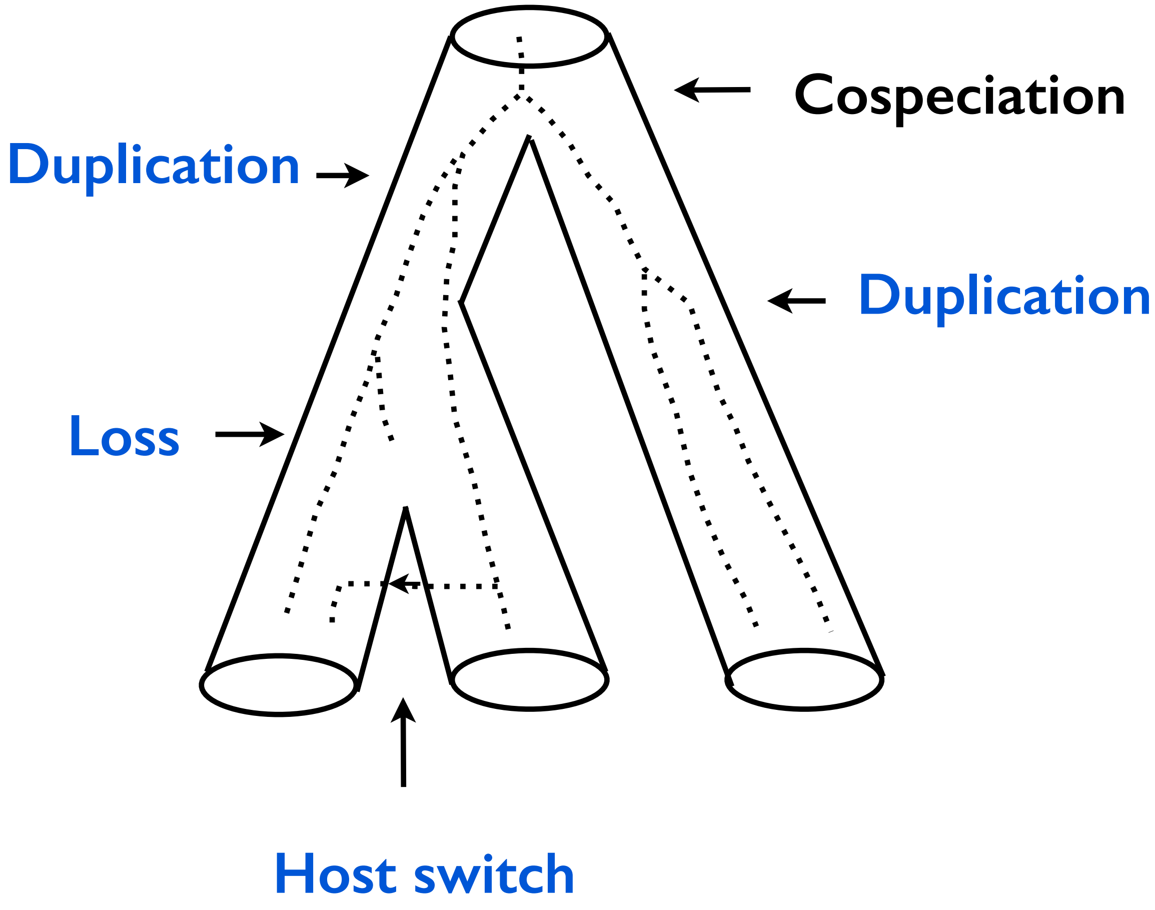
Lice



*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


Cospeciation

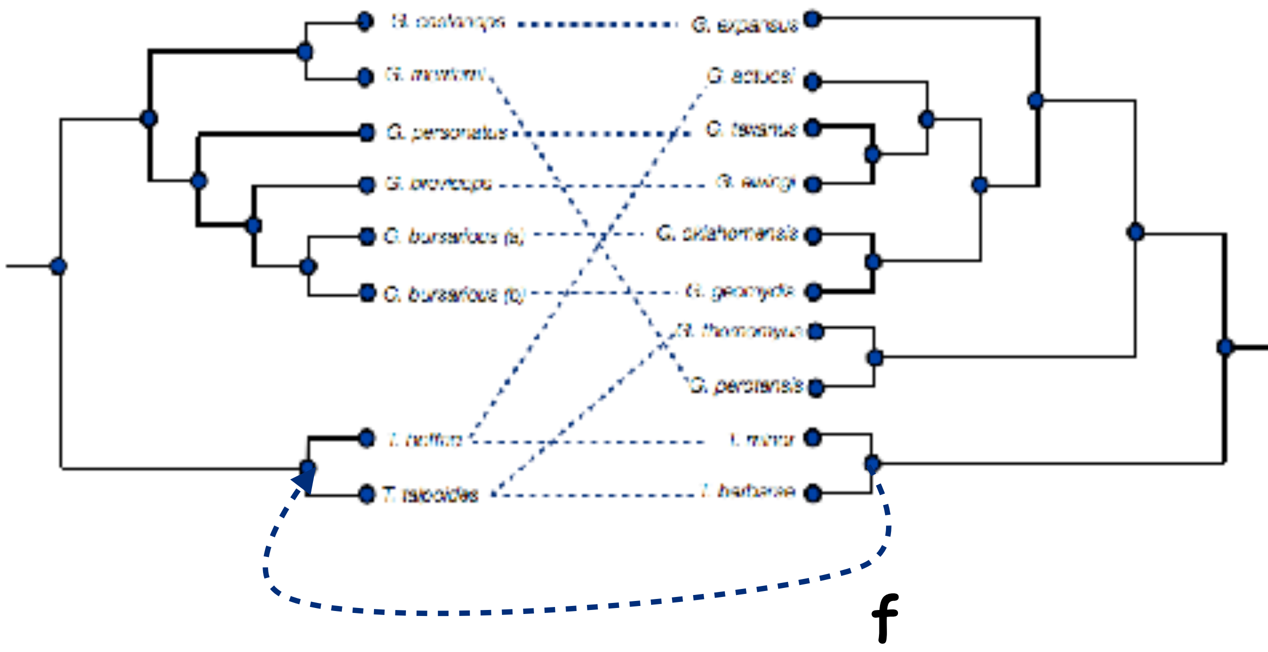
Duplication

Duplication

Loss

Host switch

- mapping/reconciliation **f**



Modeling the events

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

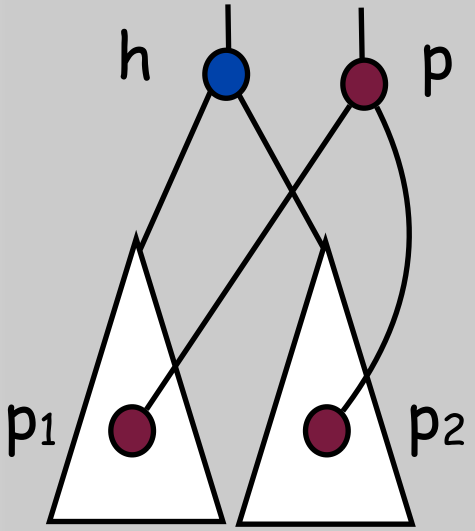
- Σ → co-speciations
- Δ → duplications
- Θ → host-switches

Modeling the events

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

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

- Co-speciation



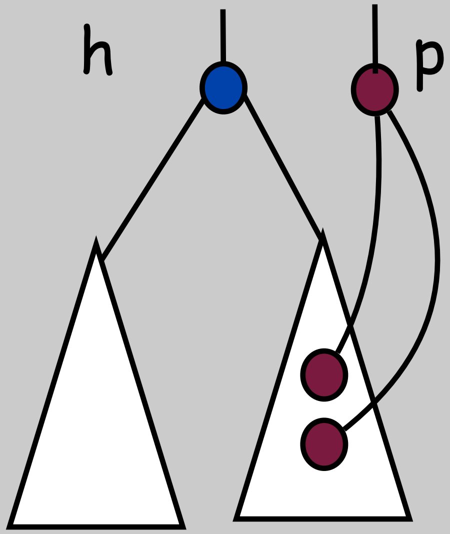
$lca(f(p_1), f(p_2))= f(p)$ and $f(p_1)$ and $f(p_2)$ are incomparable.

Modeling the events

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

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

- Duplication



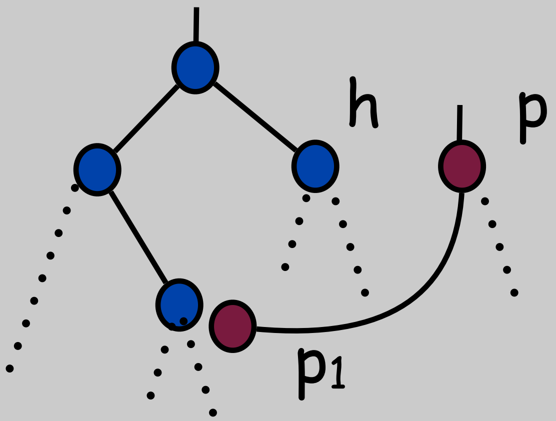
$\mathrm{lca}\ (f(p_1), f(p_2)) \in \{f(p_1),f(p_2)\}$

Modeling the events

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

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

- Host-switch

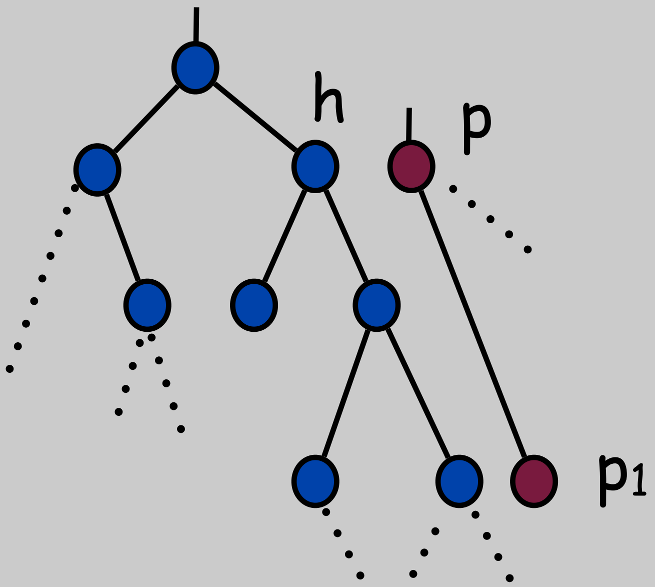


$lca(f(p_1), f(p)) \neq f(p)$

Modeling the events

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

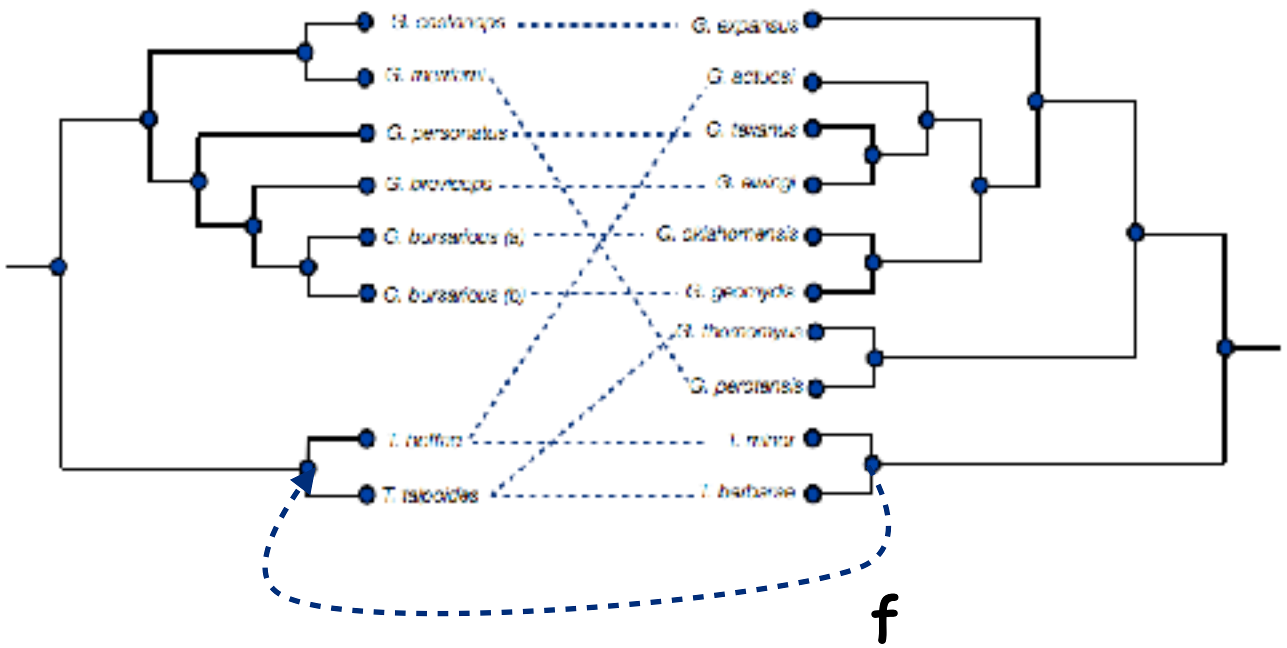
- Loss



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

Reconciliation method

Co-phylogeny reconstruction problem

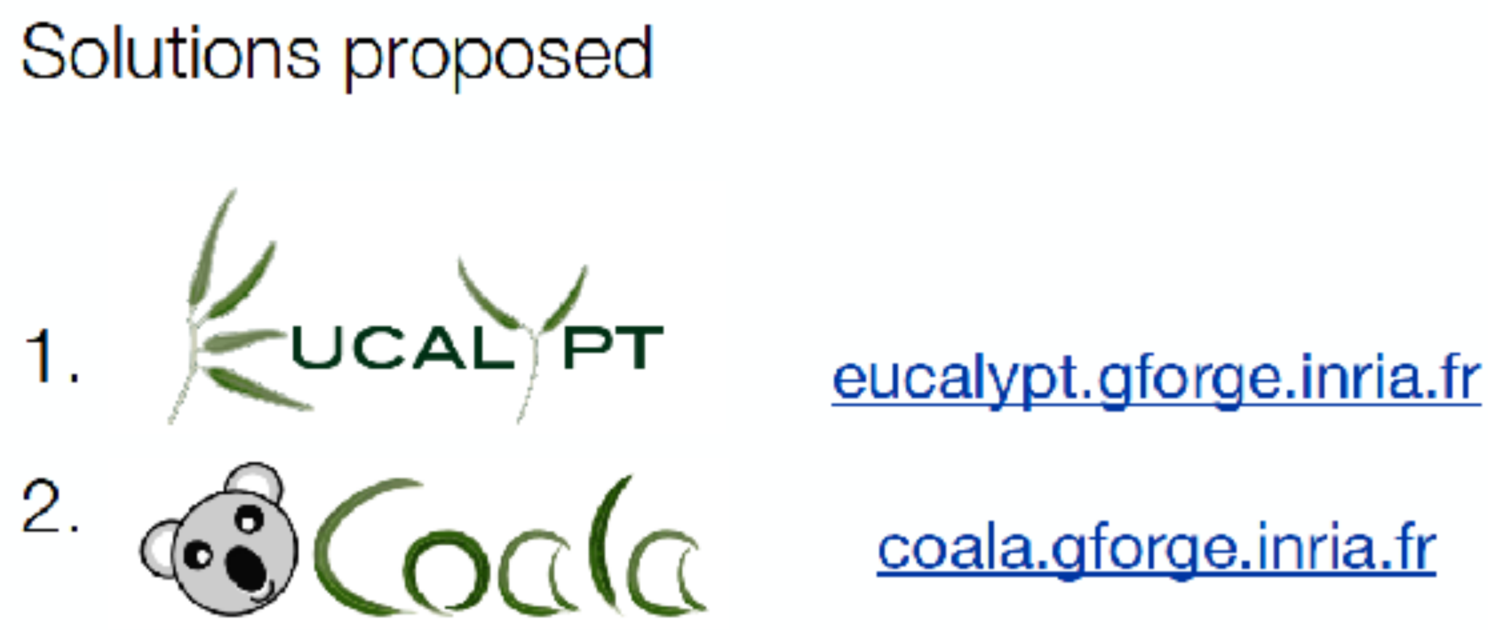


Optimality of the solution: assigns a cost to each of the four types of events and then minimizes the total cost.

Our contribution

- The cost of each of the events influences the solutions obtained. So how to choose the costs?

- Generate all optimal solutions



Solutions proposed

1. EUCALYPT eucalypt.gforge.inria.fr
2. Coala coala.gforge.inria.fr

Our contribution so far

Inference of the events cost (ABC method)

coala.gforge.inria.fr

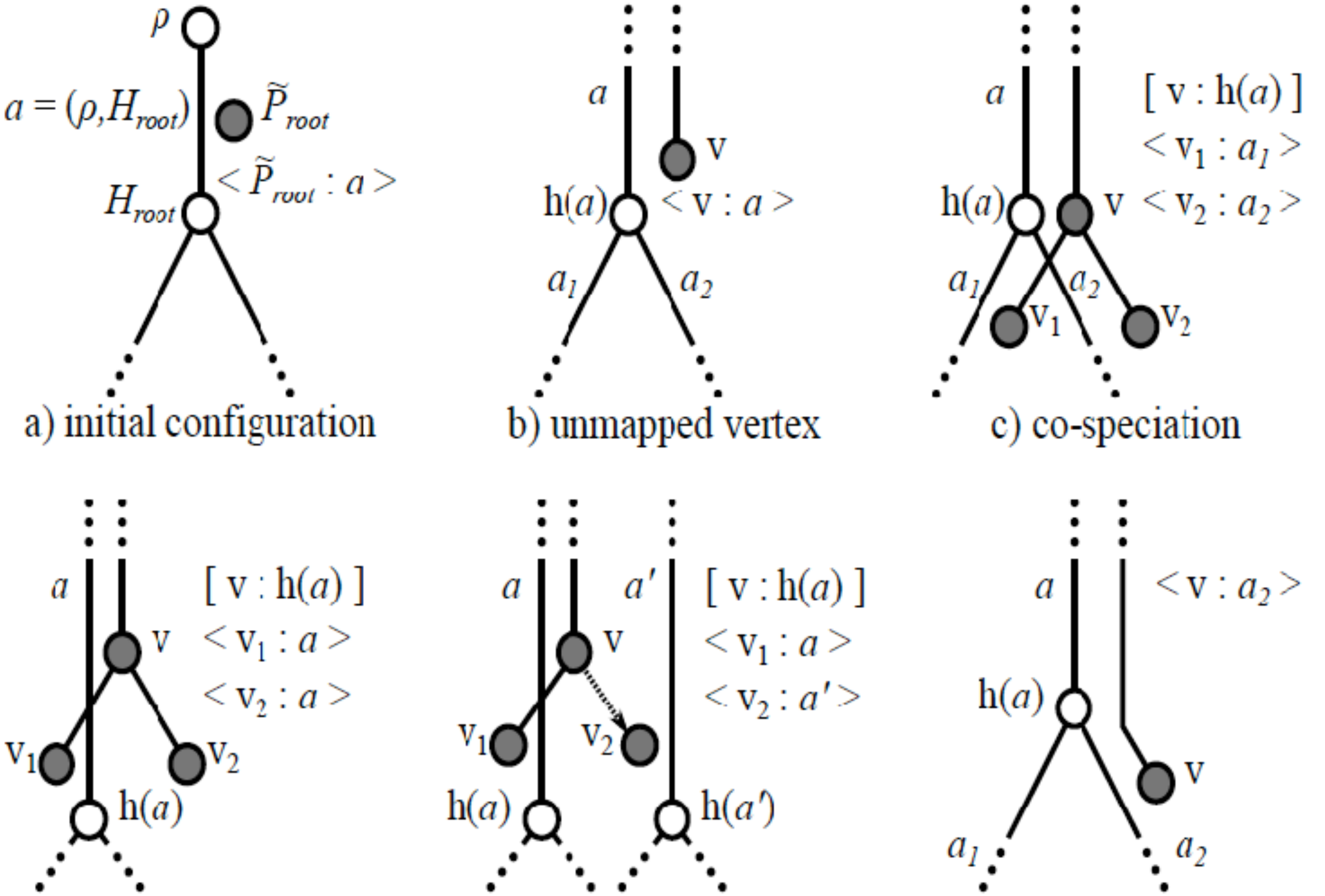
Choosing the costs

Problem: the cost of the events influences the reconciliation obtained.

Basic idea:

- For a probability vector for the events simulate the temporal evolution of the parasite tree following the evolution of the host tree.
- compare the simulated trees with the real parasite tree (using a tree metric).
- use a Sequential Monte Carlo approach to keep parameter values that generate trees “close” to the real one.

Generation of the parasite tree



a) initial configuration

b) unmapped vertex

c) co-speciation

d) duplication

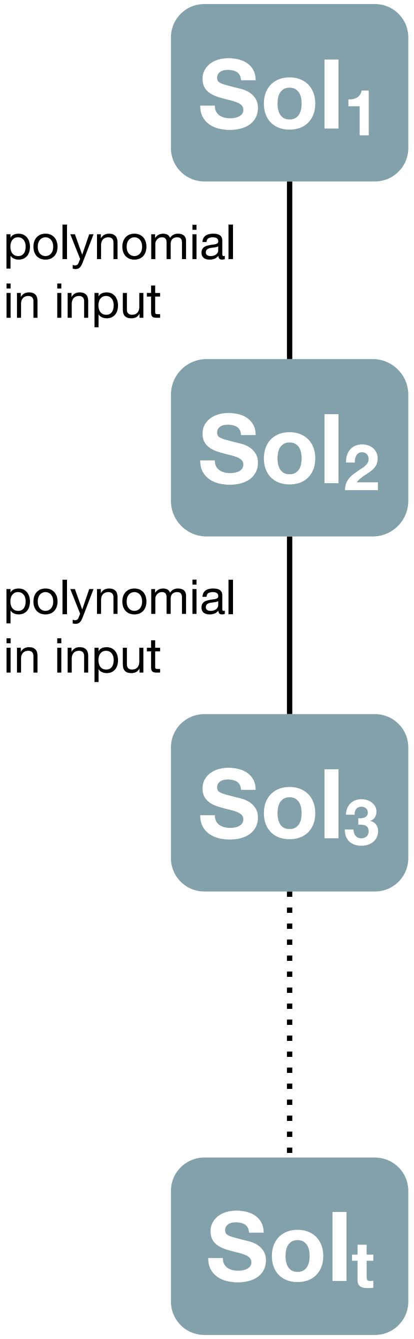
e) host-switch

f) loss

Our contribution so far

Polynomial delay enumeration algorithm

eucalypt.gforge.inria.fr


Sol_1

polynomial in input

Sol_2

polynomial in input

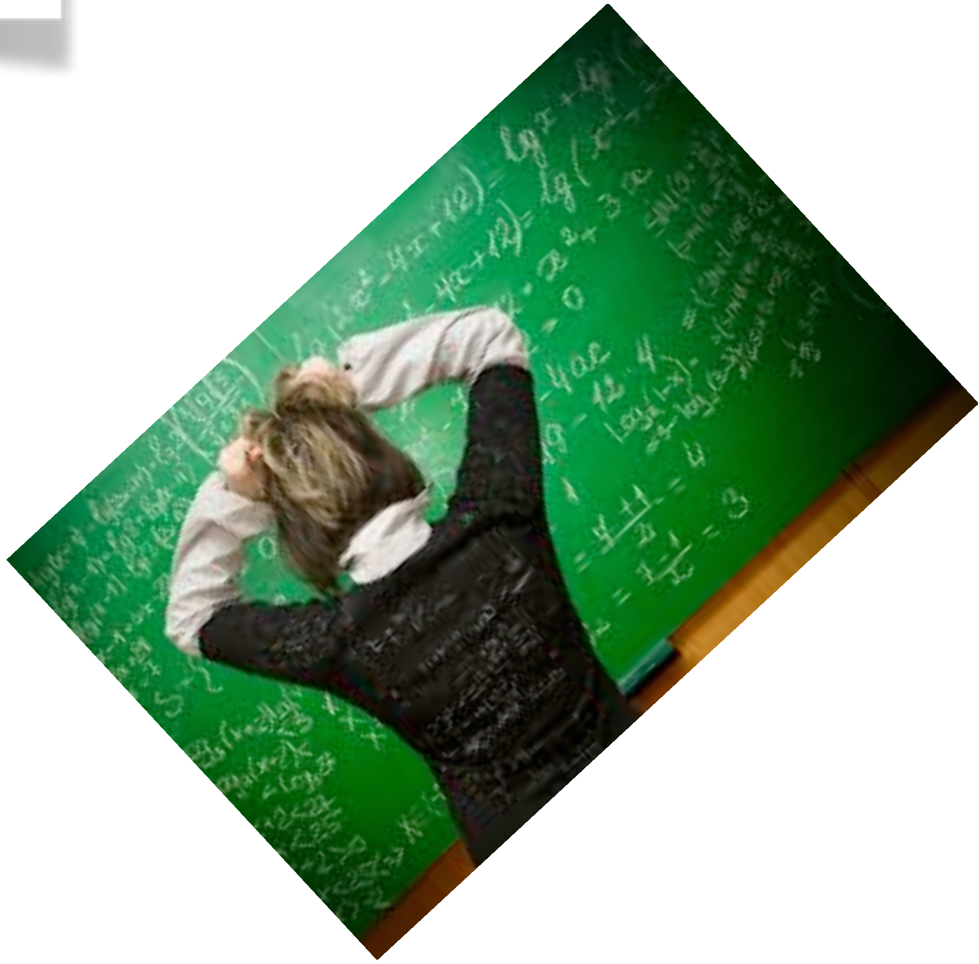
Sol_3

Sol_t

Difficulties

Generate all the optimal reconciliations.

- The number of optimal reconciliations increases rapidly even for small trees (*exponential in the size of the trees*).
- The size of the trees can be large.



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.

Problem

- Too many solutions (currently working on this)

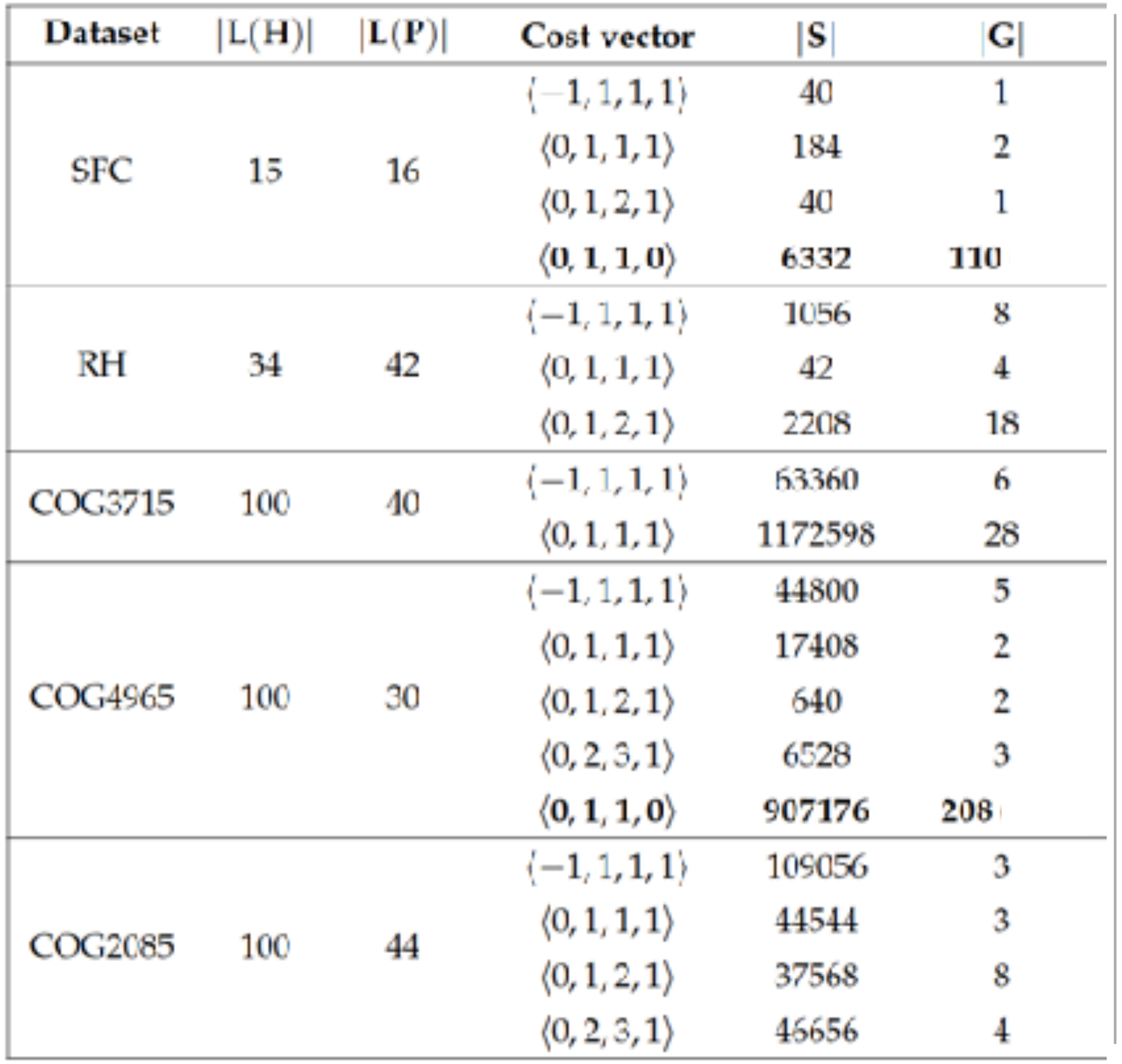
Current work

A huge number of optimal solutions:

- Group “similar" reconciliations

- Define equivalence classes
- Define a measure of similarity

Example



Dataset	|L(H)|	|L(P)|	Cost vector	|S|	|G|
SFC	15	16	$\langle -1,1,1,1\rangle$	40	1
			$\langle 0,1,1,1\rangle$	184	2
			$\langle 0,1,2,1\rangle$	40	1
			$\langle \mathbf{0},\mathbf{1},\mathbf{1},\mathbf{0}\rangle$	**6332**	**110**
RH	34	42	$\langle -1,1,1,1\rangle$	1056	8
			$\langle 0,1,1,1\rangle$	42	4
			$\langle 0,1,2,1\rangle$	2208	18
COG3715	100	40	$\langle -1,1,1,1\rangle$	63360	6
			$\langle 0,1,1,1\rangle$	1172598	28
COG4965	100	30	$\langle -1,1,1,1\rangle$	44800	5
			$\langle 0,1,1,1\rangle$	17408	2
			$\langle 0,1,2,1\rangle$	640	2
			$\langle 0,2,3,1\rangle$	6528	3
			$\langle \mathbf{0},\mathbf{1},\mathbf{1},\mathbf{0}\rangle$	**907176**	**208**
COG2085	100	44	$\langle -1,1,1,1\rangle$	109056	3
			$\langle 0,1,1,1\rangle$	44544	3
			$\langle 0,1,2,1\rangle$	37568	8
			$\langle 0,2,3,1\rangle$	46656	4

Current and future work

A huge number of optimal solutions:

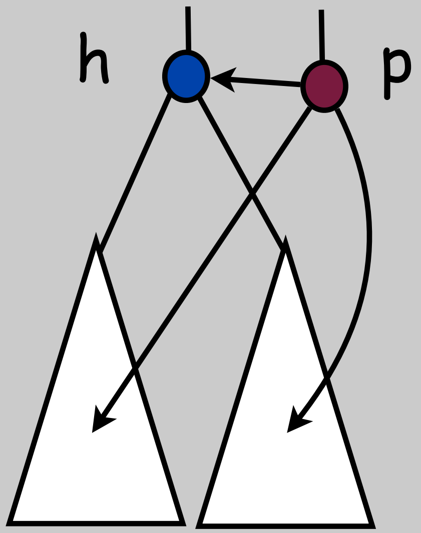
- Define a similarity measure between the reconciliations and group the solutions according to it.
- Add more biological constraints to the problem

More realistic models

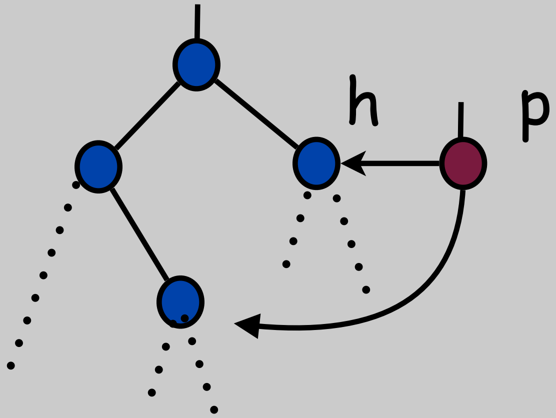
- deal with errors in the phylogenetic trees (starting from sequences)
- multiple hosts - multiple parasites (add new biological events)

New biological events

- Failure to diverge



- Spread



Questions?

http://coala.gforge.inria.fr/

http://eucalypt.gforge.inria.fr/