# Coala: CO-evolution Assessment by a Likelihood-free Approach

C. Baudet, B. Donati, B. Sinaimeri, P. Crescenzi, C. Gautier, C. Matias and M-F. Sagot

CNRS, Laboratoire Statistique et Génome, Évry
(soon Lab. Probabilités et Modèles aléatoires, Paris)
http://stat.genopole.cnrs.fr/∼cmatias





# Outline

# Context: Co-evolution and Co-phylogeny

- Co-evolution is the study of ancient relationships among ecologically linked groups of organisms, *e.g* hosts and parasites.
- Historical associations among genes, organisms and geographical areas share fundamental similarities.

## 2 parallel situations

- Hosts/Parasites systems
  - Parasites co-speciate with hosts,
  - They also independently speciate, or undergo hosts switches and "losses".
- Species/Genes evolution differ, in particular because of
  - gene duplication,
  - losses,
  - horizontal transfers.

# Reconciliation

## Definition

A reconciliation is a mapping between two trees (species/genes or hosts/parasites), with associated leaves, that maps the internal nodes of the genes (resp. parasites) tree to the internal nodes of the species (resp. hosts) tree.

## Goals and applications

- Explain divergences between phylogenetic trees of species/genes or hosts/parasites systems.
- Modeling co-evolution of these systems.
- Reconstruct species trees from (discordant) genes trees.
- ...

## Methods

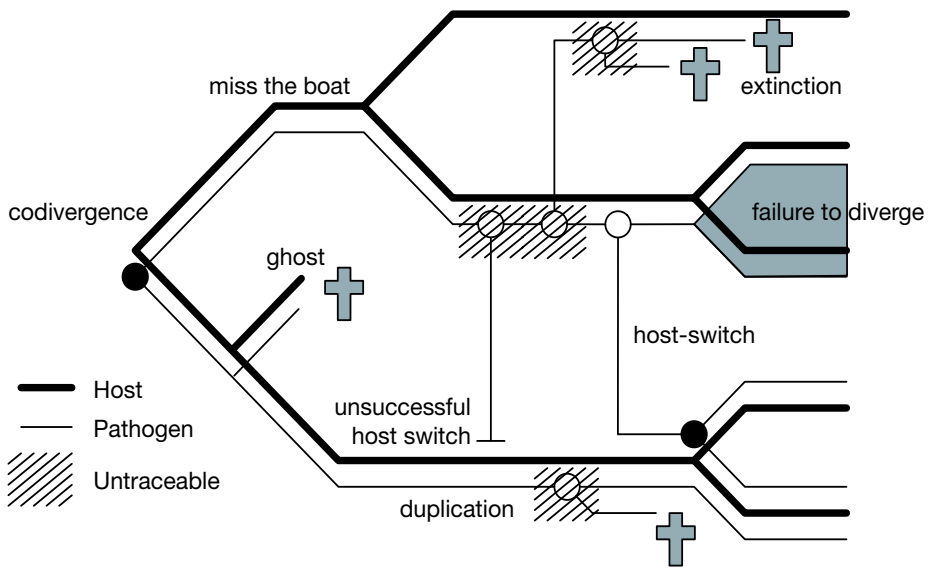
- Parsimony based methods (underestimate number of events).
- Model-based methods (preferable).

# Modeling co-evolution

## Species/Genes or Hosts/Parasites terminology

| **genes** | **parasites** |
|---|---|
| codivergence | cospeciation |
| duplication | independent speciation |
| horizontal transfer | host switch |
| loss (drift) | extinction |

## Possible vs untraceable events (source: Charleston)

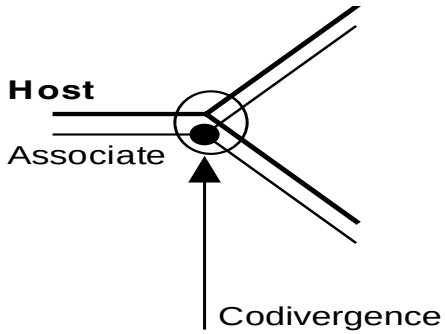


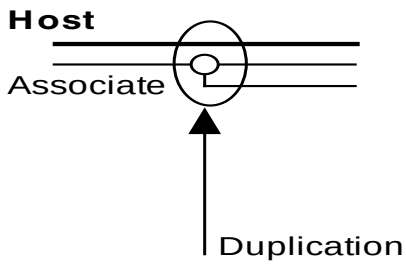

"Loss" may occur under different and undistinguishable situations

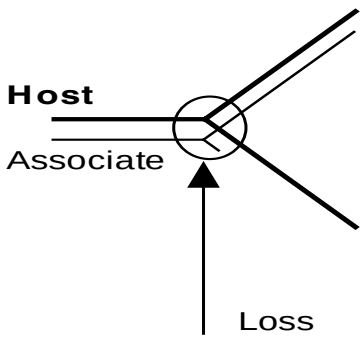
- extinction of parasite,
- failure to track both hosts after host divergence,
- sampling error.

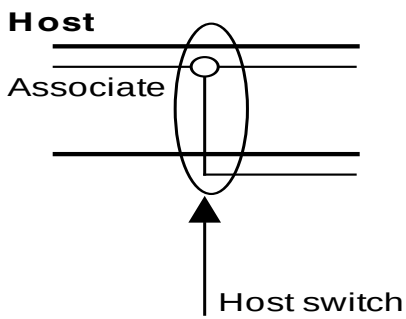
# DTL (duplication-transfer-loss) models

- Many models/methods only deal with either duplication/loss or with horizontal transfer.
- Few take into account the three type of events.
- We consider the four following co-evolution events:



# Limits of existing methods

1. Existing constraints on feasible phylogenies:
   - Hosts switches should occur only between co-existing species. If timing information is available, dynamic programming solutions exist for reconciliation.
   - Reconciliation is NP-hard when switches allowed and timing info not available.
2. Almost all methods *a priori* assign a cost to each event: crucial impact on the results !
   - Reasonable cost values are difficult to estimate
   - Different pairs of hosts/parasites phylogenies may require different event costs.
   - Exploring the space of all possible reconciliations is not feasible, thus likelihood-based approach are far from reach. $\rightarrow$ ABC procedures might be a solution.
3. ...

Here we shall deal with the second point: estimate events costs from data.

# Outline

# General overview of the method I

- The goal is to consider a co-evolution model and estimate its parameter values for a pair of hosts/parasites trees.
- Without maximising a likelihood (likelihood-free approach).
- We rely on a parasite tree generation algorithm:
  - input is a hosts tree + parameter value for co-evolution model (with four co-evolution events)
  - output is a putative parasites tree that co-evolved with hosts according to given model
- Then use approximate Bayesian inference (ABC).

# General overview of the method II

## ABC principle

Starting from an observed parasite tree (data $D_0$), iterate

- Sample a parameter value $\theta = (p_c, p_d, p_s, p_l)$ from prior $\pi$
- Generate dataset $D_\theta$ from this parameter value (following parametric model above and using hosts tree $H$),
- Compute discrepancy between $D_\theta$ and $D_0$: $d(D_0, D_\theta)$. (This may be done through a distance between summary statistics of the data).

Keep $\tau\%$ of values $\theta$ giving rise to smallest discrepancies.

ABC method belongs to the class of rejection algorithms and approximates the posterior $\mathbb{P}(\theta | d(D_0, D_\theta) \leq \epsilon)$, where $\epsilon$ is a tolerance threshold.

# Results

## What I don't tell you about

- The details about the parasites tree generation algorithm;
- The choice of prior $\pi$ and of discrepancy $d$;
- The details of the ABC procedure that we used (ABC-SMC).

## What you should trust

- The method works quite well on simulated datasets;
- It gives interesting results on real data also.

# Conclusions

## What we have done so far

- Method for estimating co-evolution parameters,
- We use these parameters for doing reconciliation with induced costs,
- Validated on synthetic data and with interesting results on real datasets

## Many remaining issues

- Refining the model, in particular towards identifiability issues,
- Enable mapping of many hosts to same parasite,
- Handle unresolved trees, weights on trees, ...
- Directly start from the sequences, not from the trees,
- ...

Thank you for listening !