

INFERRING SCENARIOS FOR  
**GENE SYNTENY–SPECIES** COEVOLUTION  
THROUGH **SEGMENTAL DUPLICATIONS, TRANSFERS,  
CUTS, GAINS AND LOSSES**

Mattéo Delabre<sup>1</sup>   Nadia El-Mabrouk<sup>1</sup>   Yoann Anselmetti<sup>2</sup>

<sup>1</sup>LBIT, Université de Montréal, Canada

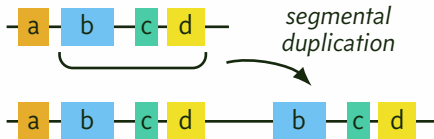
<sup>2</sup>CoBIUS Lab, Université de Sherbrooke, Canada

---

ISMB 2023 · EvolCompGen Track  
Lyon, France · July 26, 2023

## EVOLUTION OF GENE SYNTENIES

- ▶ Syntenies: groups of genes evolving together through **segmental events** (e.g., Cas systems, operons, ...)

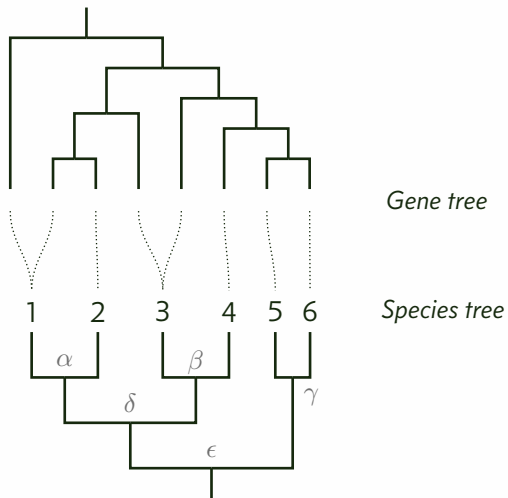


- ▶ Also **coevolving** with their host species, similarly to individual genes

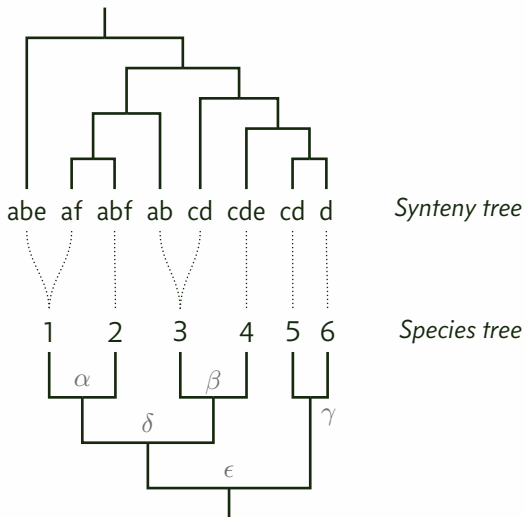
## (SUPER)-RECONCILIATIONS

- ▶ The gene–species evolutionary relationship creates **inconsistencies** between both phylogenies
- ▶ Both can be **reconciled** by postulating past events (duplications, lateral transfers, ...)
- ▶ These events form candidate **coevolution scenarios**
- ▶ **Super-reconciliation** extends this model to segmental events that describe **synteny evolution**

# EXAMPLE

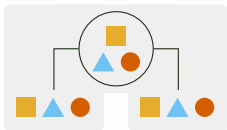


# EXAMPLE

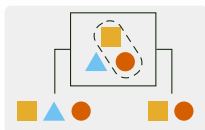


## EVOLUTION MODEL

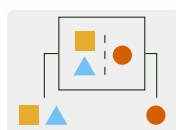
- ▶ Syntenies are **sets of genes** (*disregards gene orders and multiplicity*)
- ▶ **Divergence events** create new synteny copies
- ▶ **Unary events** change existing syntenies



Speciation



Duplication



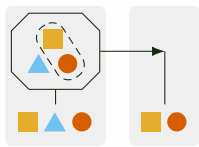
Cut



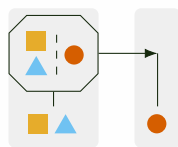
Gain



Loss



Duplication with transfer



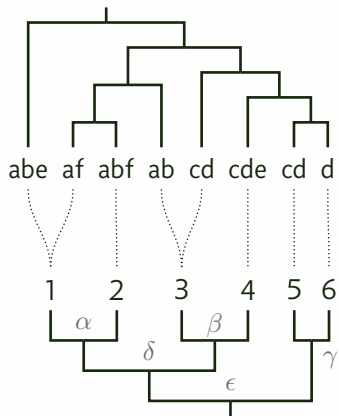
Cut with transfer

## SCENARIOS AND MISSING SPECIES

- ▶ Scenario: tree of events
- ▶ **Leaves** and **topology** match those of the studied synteny tree
- ▶ Events happen along the specified species tree
- ▶ **Unsampled** or **extinct** species can be postulated

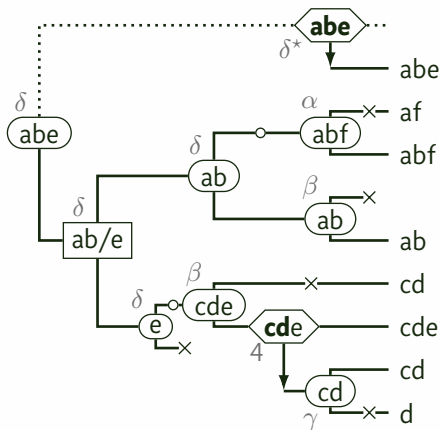
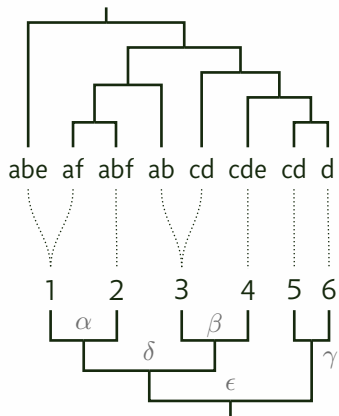
*Weiner and Bansal. "Improved duplication-transfer-loss reconciliation with extinct and unsampled lineages" (Aug. 2021)*

# EXAMPLE





# EXAMPLE



## LOOKING FOR MAXIMUM PARSIMONY

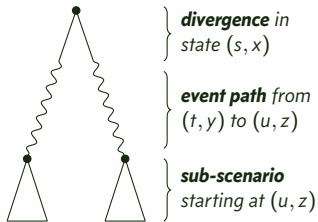
- ▶ Each event type is given a **cost** ( $\sim$  inverse probability)
- ▶ Find the **most parsimonious** (least costly) scenarios for a given input synteny and species tree
  - *For example, the previous scenario is optimal if losses cost 1, duplications/cuts cost 2, and transfers cost 3*

## LOOKING FOR MAXIMUM PARSIMONY

- ▶ Each event type is given a **cost** ( $\sim$  inverse probability)
- ▶ Find the **most parsimonious** (least costly) scenarios for a given input synteny and species tree
  - *For example, the previous scenario is optimal if losses cost 1, duplications/cuts cost 2, and transfers cost 3*
  
- ▶ Space of possible scenarios is **unbounded** in size
- ▶ Number of optimal scenarios can be **exponential**

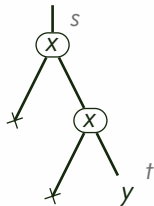
## BUILDING SCENARIOS WITH EVENT PATHS

- ▶ Valid scenarios **pass through all divergences** of the synteny tree
- ▶ Scenarios can be decomposed into **sub-scenarios** starting at each divergence, linked by **event paths**
- ▶ Event paths start at a given state and have a single **visible** leaf: the target state

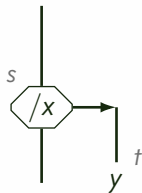


## POSSIBLE OPTIMAL EVENT PATHS

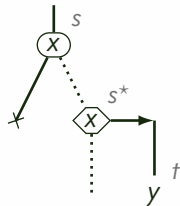
- ▶ The number of possible optimal event paths leading from  $(s, x)$  to  $(t, y)$  is bounded by a **constant**



*Chain of speciations  
and losses*



*Full cut with transfer  
(s separate from t)*



*Transfer through  
unsampled species*

## DYNAMIC PROGRAMMING FORMULATION

- ▶  $c(v, s, x)$ : least cost of scenarios...
  - for the syntenic tree **below**  $v$
  - where  $v$ 's **host** is  $s$
  - and its **contents** are  $x$
- ▶ Compute  $c(v, s, x)$  recursively by trying **all possible event types**
  - *Example for a speciation:*

$$c(v, s, x) = \min_{\substack{u \leq s_\ell \\ w \leq s_r \\ y, z}} \{ c(P[(s_\ell, x) \rightarrow (u, y)]) + c(v_\ell, u, y) + c(P[(s_r, x) \rightarrow (w, z)]) + c(v_r, w, z) \}$$

## DYNAMIC PROGRAMMING FORMULATION

- ▶  $c(v, s, x)$ : least cost of scenarios...
  - for the syntenic tree **below**  $v$
  - where  $v$ 's **host** is  $s$
  - and its **contents** are  $x$
- ▶ Compute  $c(v, s, x)$  recursively by trying **all possible event types**
  - *Example for a speciation:*

$$c(v, s, x) = \min_{\substack{u \leq s_\ell \\ w \leq s_r \\ y, z}} \{ c(P[(s_\ell, x) \rightarrow (u, y)]) + c(v_\ell, u, y) + c(P[(s_r, x) \rightarrow (w, z)]) + c(v_r, w, z) \}$$

- ▶ **Gains** are placed at the LCA
- ▶ Only two contents need to be tried:  
the **minimum valid set** of genes and **any bigger set**

## DYNAMIC PROGRAMMING APPLICATIONS

- ▶ **Enumerate** most parsimonious solutions
- ▶ **Count** how many optimal solutions exist

- ▶ Uniformly **sample** among optimal solutions

*Bansal, Alm, and Kellis. "Reconciliation revisited: handling multiple optima when reconciling with duplication, transfer, and loss" (Oct. 2013)*

- ▶ Enumerate **Pareto-optimal** solutions

*Libeskind-Hadas, Wu, Bansal, and Kellis. "Pareto-optimal phylogenetic tree reconciliation" (June 2014)*

- ▶ Enumerate **equivalence classes** of solutions

*Wang, Mary, Sagot, and Sinaimeri. "Efficiently sparse listing of classes of optimal cophylogeny reconciliations" (Feb. 2022)*



## ONGOING AND FUTURE WORK

- ▶ Preliminary *implementation* available at [github.com/UdeM-LBIT/superrec2](https://github.com/UdeM-LBIT/superrec2)
- ▶ Ongoing *evaluation* on simulated and real data

### Open questions:

- ▶ **Build synteny trees** from individual gene trees
- ▶ Consider **fusion** events (reconciliation of networks)  
*Chan and Robin. "Reconciliation of a gene network and species tree" (July 2019)*
- ▶ Handle **gene multiplicities**
- ▶ Minimize number of postulated unsampled species