SYNESTH: COMPREHENSIVE SYNTENIC RECONCILIATION WITH UNSAMPLED LINEAGES

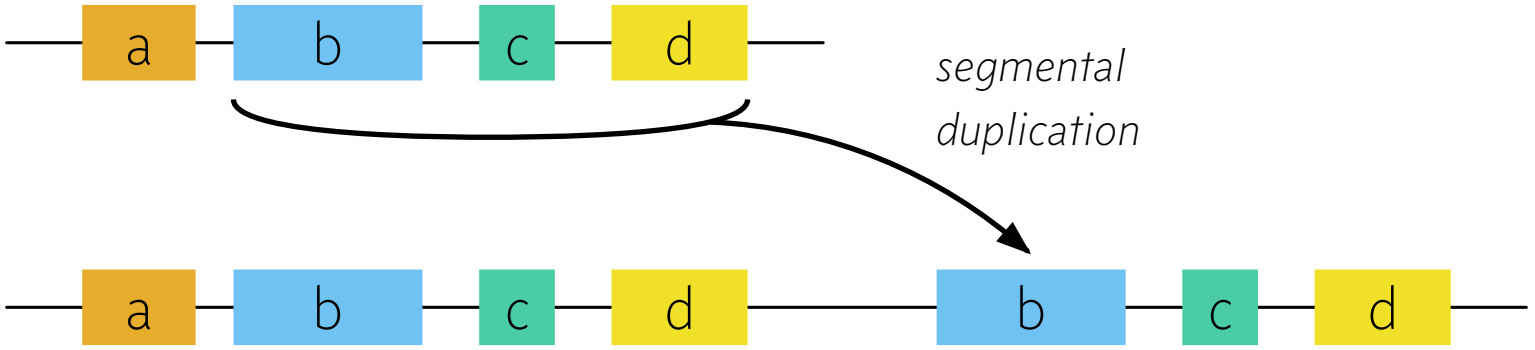
Mattéo Delabre Nadia El-Mabrouk

LBIT, DIRO, Université de Montréal, Canada

Quest for Orthologs 2024
Montréal, Canada · July 17, 2024

MOTIVATION

- **Synteny:** Group of *co-localized genes* evolving together (e.g., CRISPR-Cas systems, operons, …)

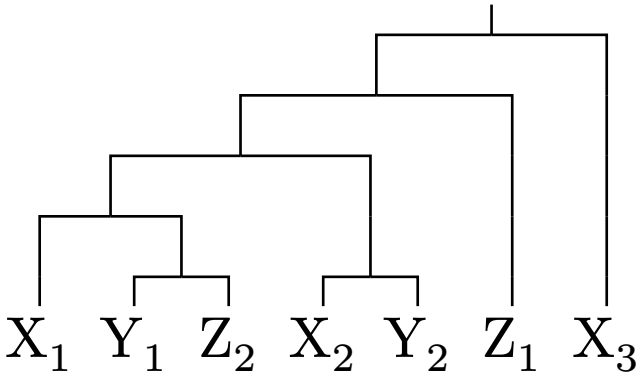


GOAL

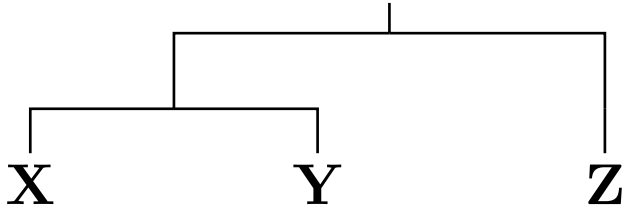
Infer the **evolutionary relationships** of homologous syntenies

RECONCILIATION

- Infer relationships of homologous **genes** by mapping nodes of a **gene tree** onto the nodes of its associated **species tree**



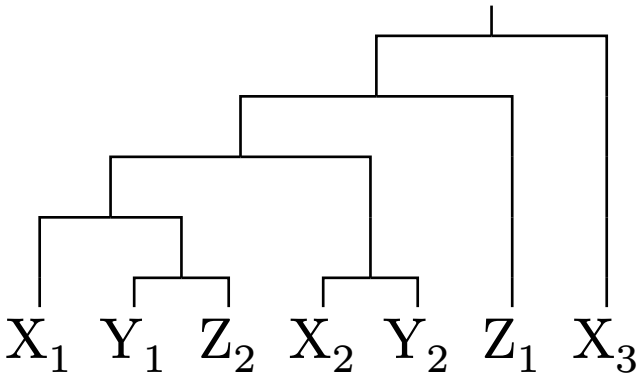
Gene tree



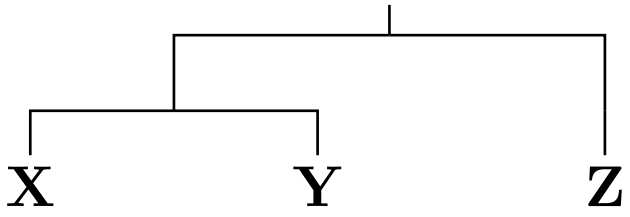
Species tree

RECONCILIATION

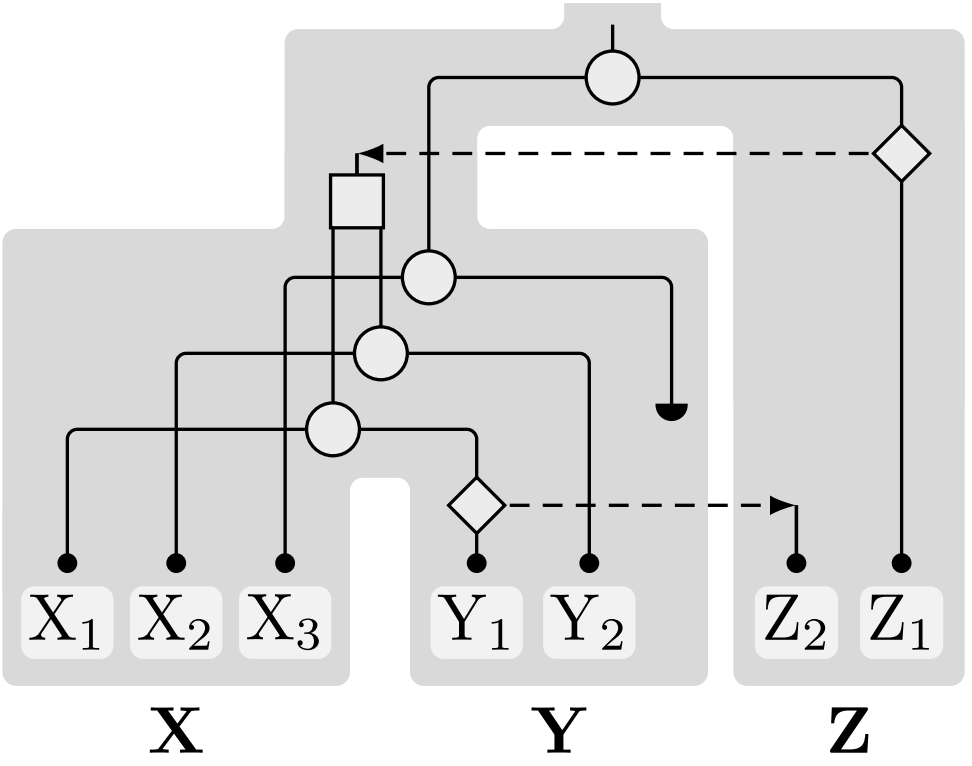
- Infer relationships of homologous **genes** by mapping nodes of a **gene tree** onto the nodes of its associated **species tree**



Gene tree



Species tree

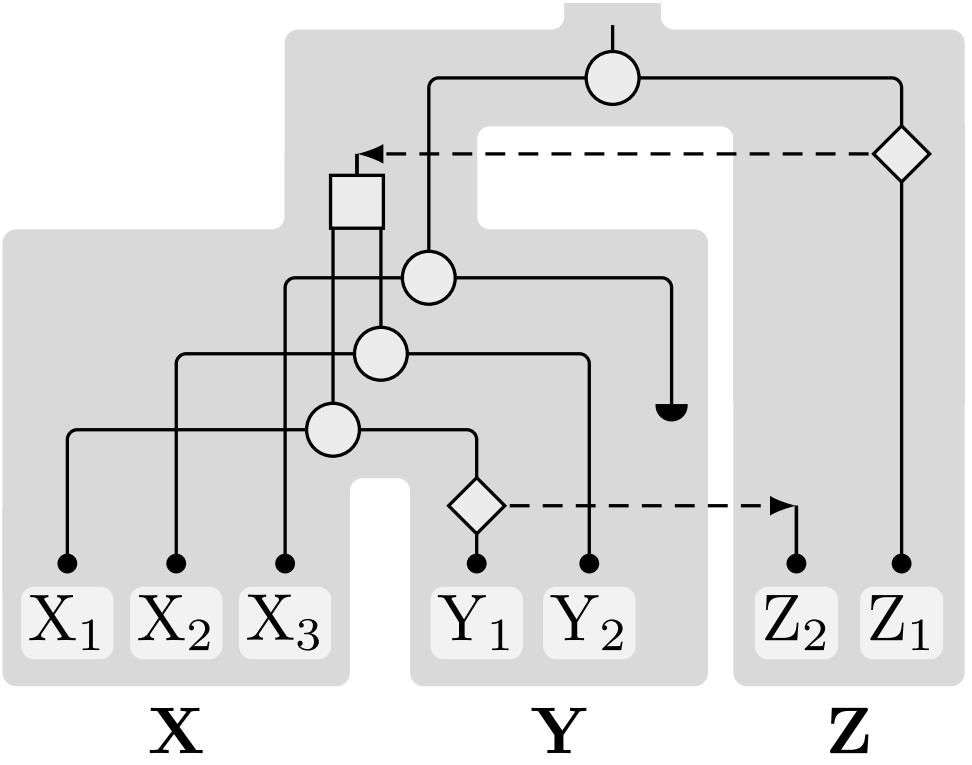


Possible reconciled history

GENE EVOLUTION MODEL

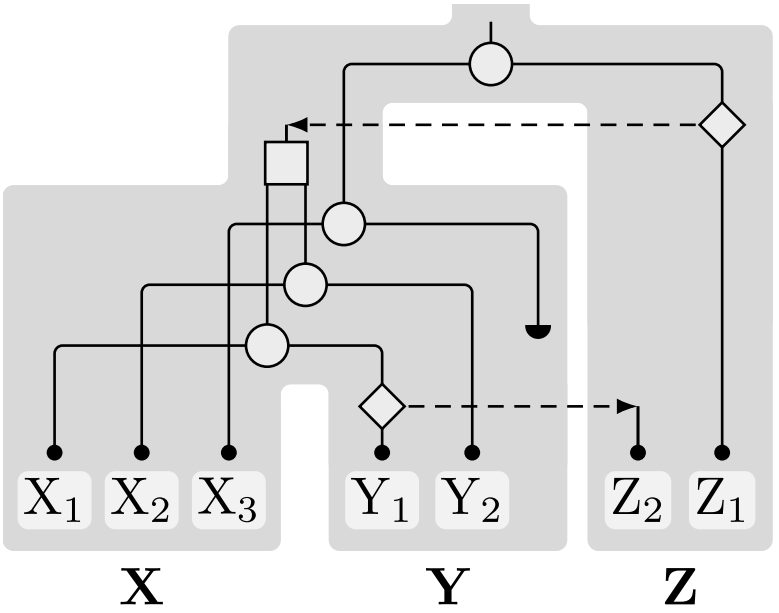
Histories are **sequences of events** chosen among:

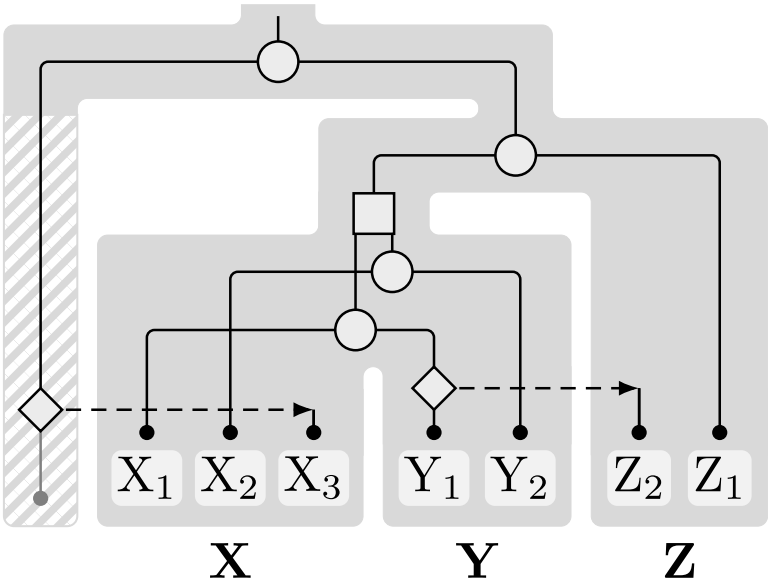
- (Co)speciations
- Duplications
- Horizontal transfers
- Losses



UNSAMPLED/EXTINCT SPECIES

- Because of transfers, histories may interact with **unknown** species
- **Allow for events** to and from those external species





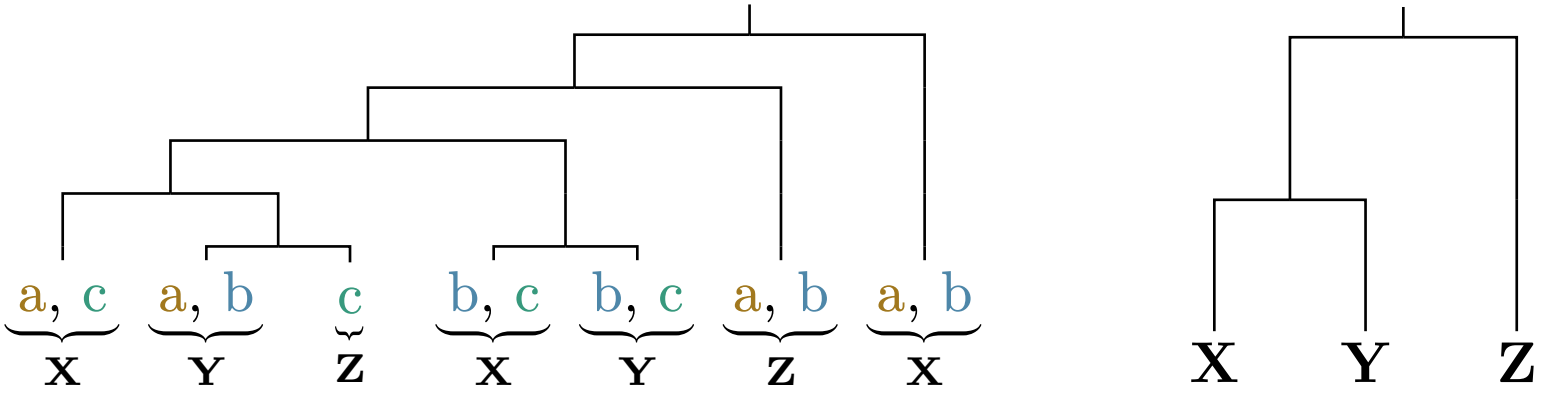
S. Weiner and M. S. Bansal, *Improved duplication-transfer-loss reconciliation with extinct and unsampled lineages,* August 2021 (Algorithms)

SYNTENIC RECONCILIATION

OUR CONTRIBUTION

Extend the reconciliation approach to **synteny trees**

- A **synteny tree** is a tree on **sets of gene families** (without orders)



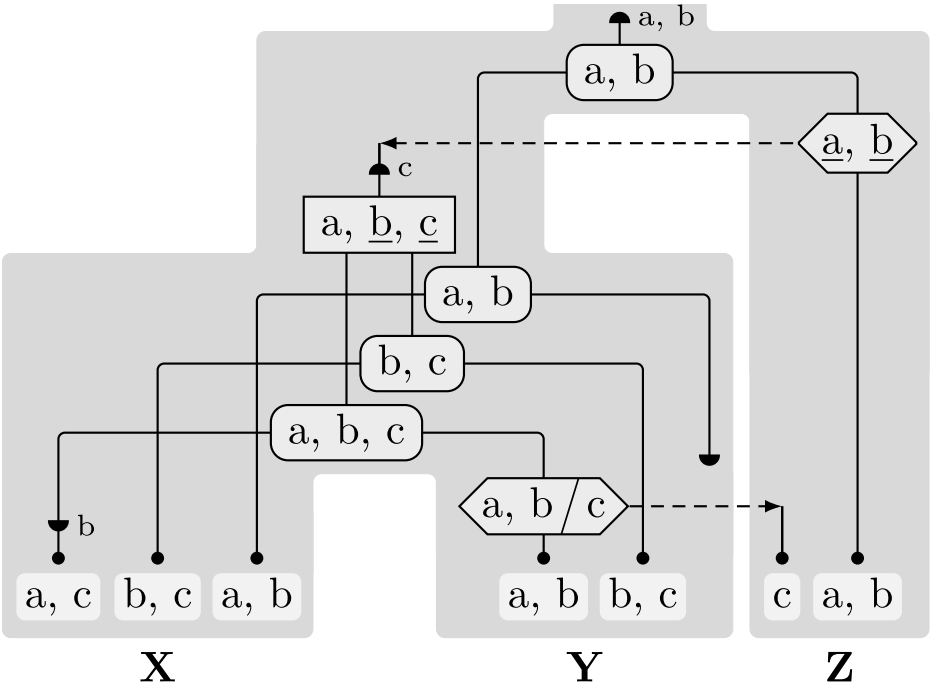
Synteny tree *Species tree*

- *Reduces to a gene tree if all sets contain the same single family*

SYNTENY EVOLUTION MODEL

Events act on **sets of genes**:

- a, b — Gains
- a, b — Losses
- a, b — (Co)speciations
- a, b — Duplications
- a, b — Transfers
- a / b — Cuts
- a / b — *Cutting* transfers



SEARCHING FOR PLAUSIBLE HISTORIES

A cost-based **parsimony** framework is generally used:

- Assign a fixed **cost** to each type of event
- Cost of a history is the **sum of costs** of each event it contains
- Look for histories of **minimum total cost**

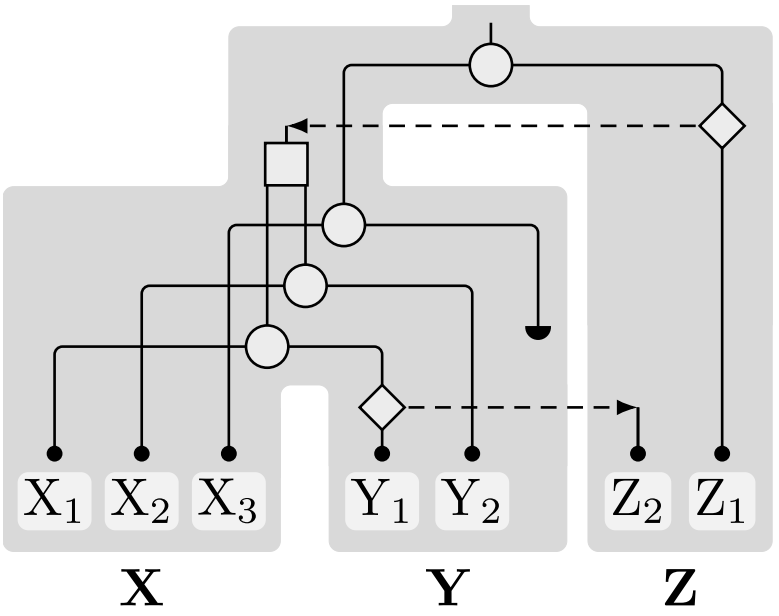
COMPUTATIONAL PROBLEM

Input: Gene (or synteny) tree; Species tree; Set of costs
Output: Minimum-cost histories corresponding to the trees

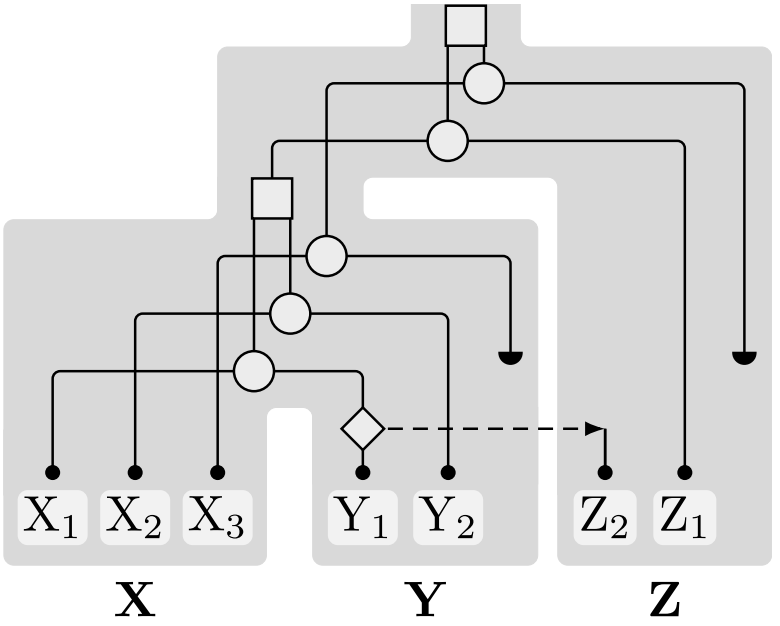
- Unfortunately, choosing the right event costs is **challenging**

A COST-FREE METHOD

- Associate each history to a **vector of number of events** of each type
- Look for histories such that **no history with a better vector** exists



$(\square\ 1, \diamond\ 2, \downarrow 1)$



$(\square\ 2, \diamond\ 1, \downarrow 2)$

PARETO-OPTIMAL HISTORIES

- Histories with non-dominated event vectors are **Pareto-optimal**

COMPUTATIONAL PROBLEM

Input: Gene (or synteny) tree; Species tree; ~~Set of costs~~
Output: Pareto-optimal histories corresponding to the trees

- *Eqv.:* Histories for which there exists costs making them optimal
- Gives an **overview** of the outcomes of all possible cost settings

---

R. Libeskind-Hadas, Y.-C. Wu, M. S. Bansal, M. Kellis, *Pareto-optimal phylogenetic tree reconciliation,* June 2014 (Bioinformatics)

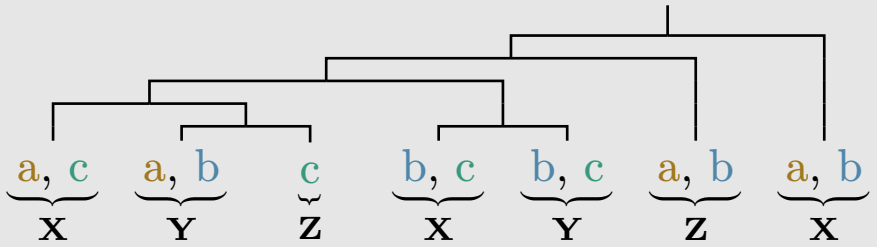
SOLVING SYNTENIC RECONCILIATION

- **Algebraic dynamic programming** approach
- We devise **recurrence relations** generating possible histories
  - That set is technically **infinite**
  - **Only a finite subset** can ever be Pareto-optimal
  - *(Tedious case analysis because of the large set of events)*
- We adapt **semiring** structures to compute various desired results

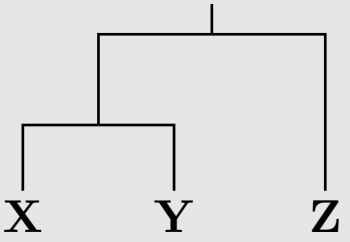
---

M. Delabre and N. El-Mabrouk, *Synesth: Comprehensive syntenic reconciliation with unsampled lineages,* April 2024 (Algorithms)

WHAT CAN WE COMPUTE?



Synteny tree (T)

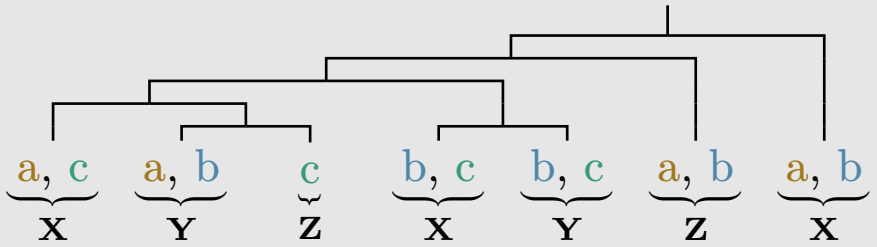


Species tree (S)

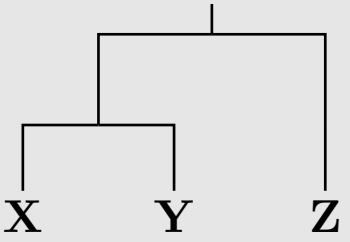
Total number of histories: 25,457,672,160,297

- *Time:* $\mathcal{O}(|V(T)|\ |V(S)|^2)$

WHAT CAN WE COMPUTE?



Synteny tree (T)

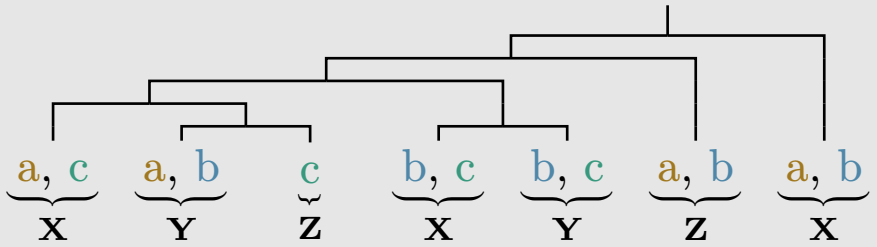


Species tree (S)

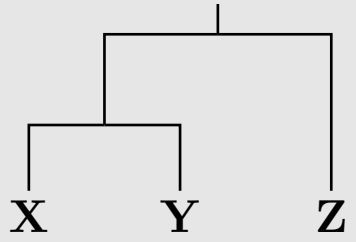
Minimum cost of any history (unit costs): 4

- *Time:* $\mathcal{O}(|V(T)|\ |V(S)|^2)$

WHAT CAN WE COMPUTE?



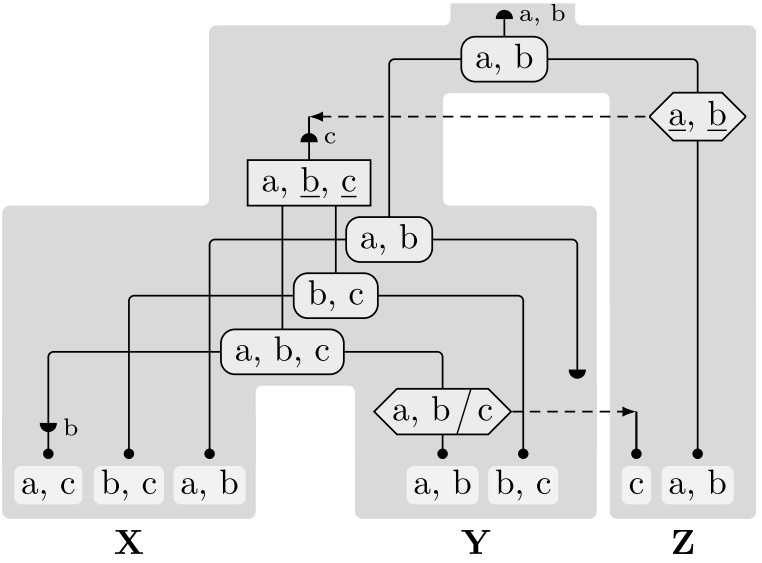
Synteny tree (T)



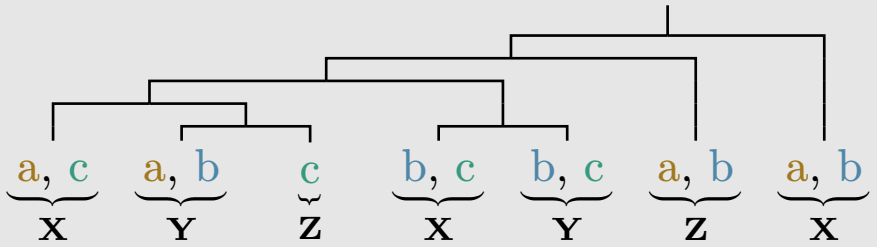
Species tree (S)

Single minimum-cost history

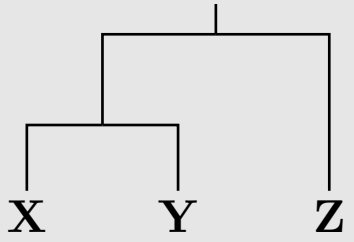
- *Time:* $\mathcal{O}(|V(T)|\ |V(S)|^3)$



WHAT CAN WE COMPUTE?



Synteny tree (T)

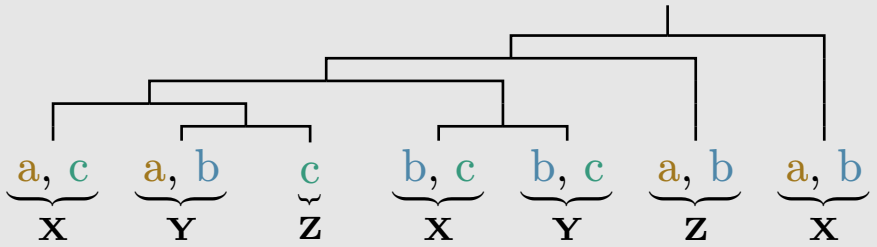


Species tree (S)

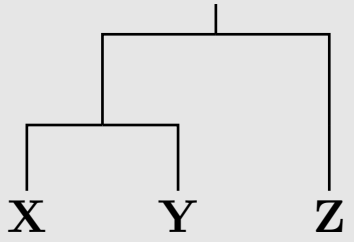
Number of co-optimal, minimum-cost, histories (unit costs): 48

- *Time:* $\mathcal{O}(|V(T)|\ |V(S)|^2)$

WHAT CAN WE COMPUTE?



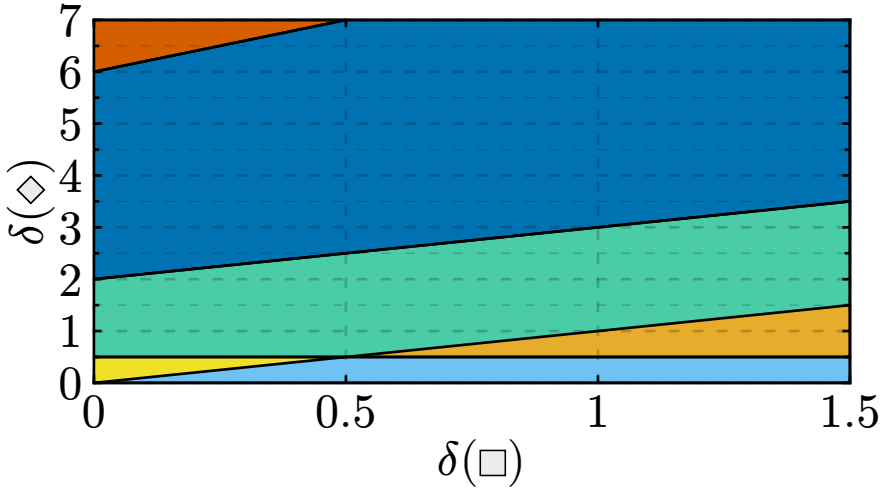
Synteny tree (T)

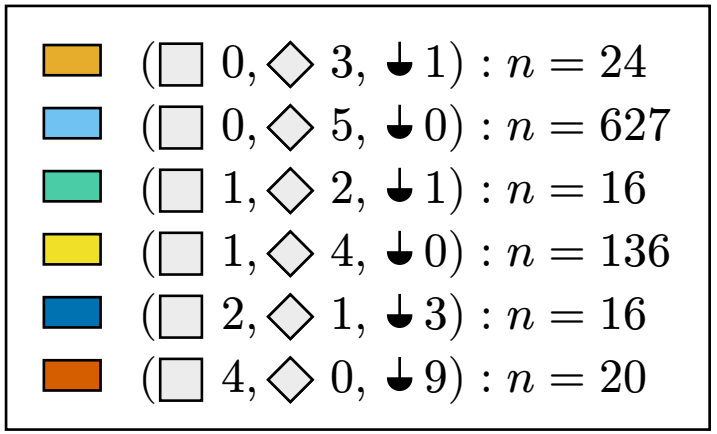


Species tree (S)

Pareto-optimal event counts

Time: $\tilde{\mathcal{O}}\left(\left(|V(T)|\ |V(S)|\right)^9\right)$





REMAINING CHALLENGES

In increasing(?) difficulty order:

- **Validate** the model on simulated and real datasets
-
-
-
-
-
-

REMAINING CHALLENGES

In increasing(?) difficulty order:

- **Validate** the model on simulated and real datasets
- Select histories that use the **least number** of **unsampled** species
-
-
-
-
-

REMAINING CHALLENGES

In increasing(?) difficulty order:

- **Validate** the model on simulated and real datasets
- Select histories that use the **least number** of **unsampled** species
- Navigate and summarize the space of **co-optimal histories**
-
-
-
-

REMAINING CHALLENGES

In increasing(?) difficulty order:

- **Validate** the model on simulated and real datasets
- Select histories that use the **least number** of **unsampled** species
- Navigate and summarize the space of **co-optimal histories**
- Deal with **time-inconsistent** optimal histories
-
-
-

REMAINING CHALLENGES

In increasing(?) difficulty order:

- **Validate** the model on simulated and real datasets
- Select histories that use the **least number** of **unsampled** species
- Navigate and summarize the space of **co-optimal histories**
- Deal with **time-inconsistent** optimal histories
- **Automate** the derivation of the dynamic programming **recurrence**
-
-

REMAINING CHALLENGES

In increasing(?) difficulty order:

- **Validate** the model on simulated and real datasets
- Select histories that use the **least number** of **unsampled** species
- Navigate and summarize the space of **co-optimal histories**
- Deal with **time-inconsistent** optimal histories
- **Automate** the derivation of the dynamic programming **recurrence**
- Incorporate **gene order** information
-

REMAINING CHALLENGES

In increasing(?) difficulty order:

- **Validate** the model on simulated and real datasets
- Select histories that use the **least number** of **unsampled** species
- Navigate and summarize the space of **co-optimal histories**
- Deal with **time-inconsistent** optimal histories
- **Automate** the derivation of the dynamic programming **recurrence**
- Incorporate **gene order** information
- Allow for **synteny fusions** and **tandem duplications**

CONCLUSION

- **Synesth** is a model and algorithm for **syntenic reconciliation**
- Infers **evolutionary relationships** between **homologous syntenies**

- Implemented in `superrec2`,
  our Python package for reconciliation
  https://github.com/UdeM-LBIT/superrec2

Thanks to the LBIT team, and to
NSERC and FRQNT for their funding support