SUPER-RECONCILIATION WITH HORIZONTAL GENE TRANSFERS

Mattéo Delabre Nadia El-Mabrouk

University of Montreal
July 28th, 2021

EVOLUTION OF SYNTENIES[1]

Definition: SYNTENY

Two or more genomic regions derived from a shared ancestral region

- Examples: Operons in bacteria, Homeobox gene clusters, …
- Should be taken into account for parsimony-based phylogeny

[1] Nadia El-Mabrouk. "Predicting the evolution of syntenies—An algorithmic review." In: *Algorithms* 14.5 (May 2021), p. 152. DOI: 10.3390/a14050152.

EVOLUTION OF SYNTENIES[1]

Definition: SYNTENY

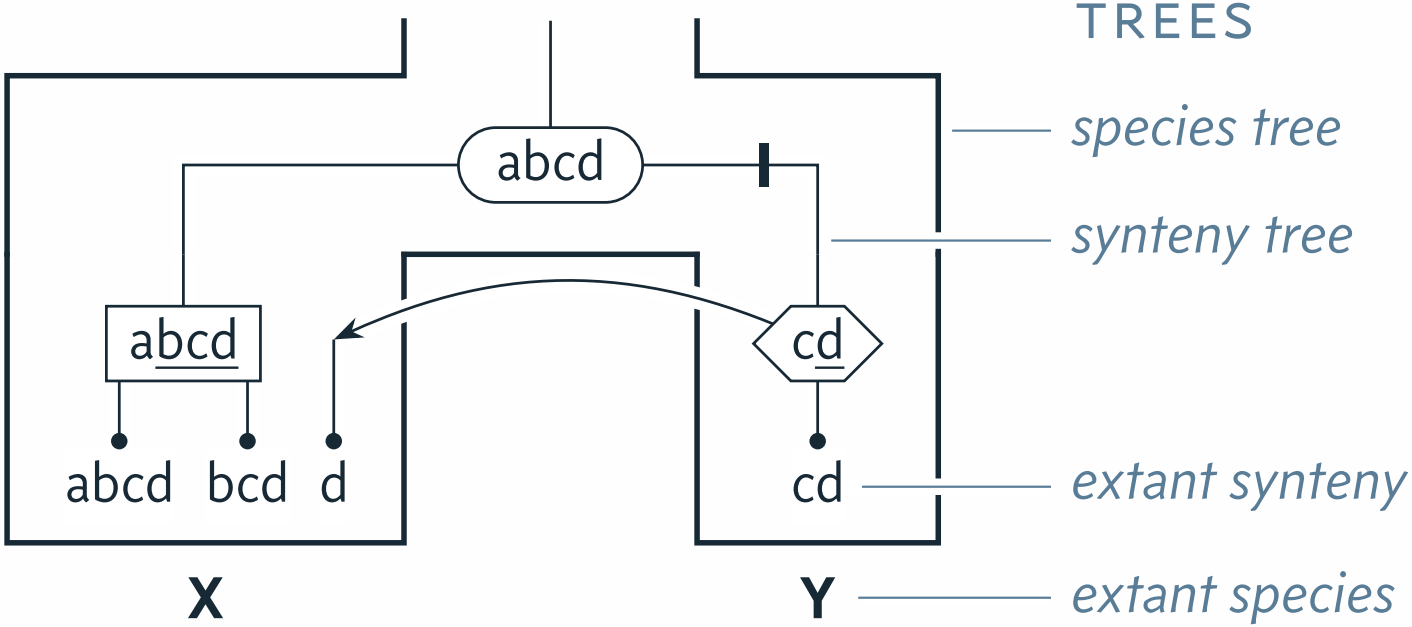
Two or more genomic regions derived from a shared ancestral region

- Examples: Operons in bacteria, Homeobox gene clusters, …
- Should be taken into account for parsimony-based phylogeny
- Existing methods:
  - DILTAG (Lajoie et al., 2010)
  - DeCoSTAR (Duchemin et al., 2017)
  - Duplication Episodes (Paszek and Górecki, 2018)
  - **Super-Reconciliation** (Delabre et al., 2020)

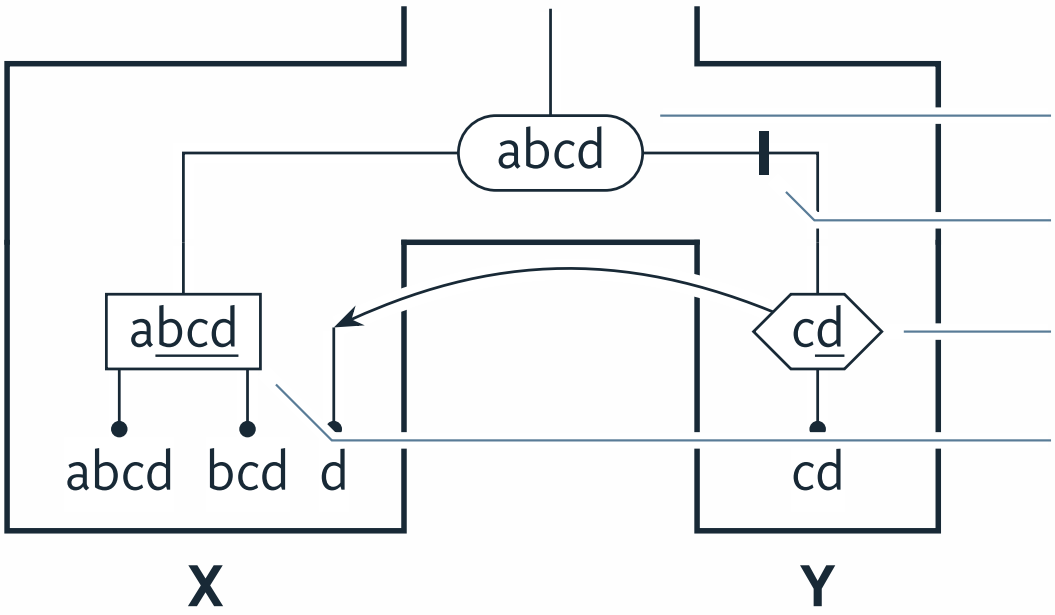
[1] Nadia El-Mabrouk. "Predicting the evolution of syntenies—An algorithmic review." In: *Algorithms* 14.5 (May 2021), p. 152. DOI: 10.3390/a14050152.

OUTLINE

ELEMENTS OF A SUPER-RECONCILIATION



ELEMENTS OF A SUPER-RECONCILIATION



EVENTS

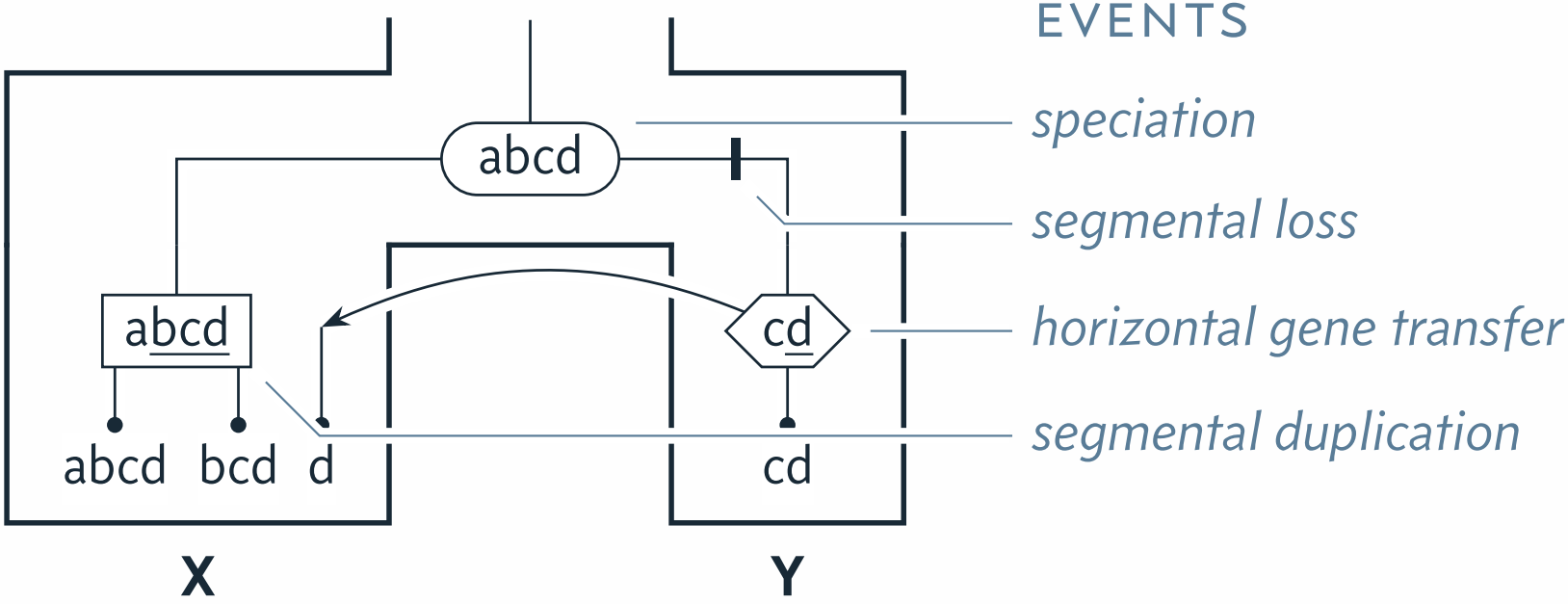
speciation

segmental loss

horizontal gene transfer

segmental duplication

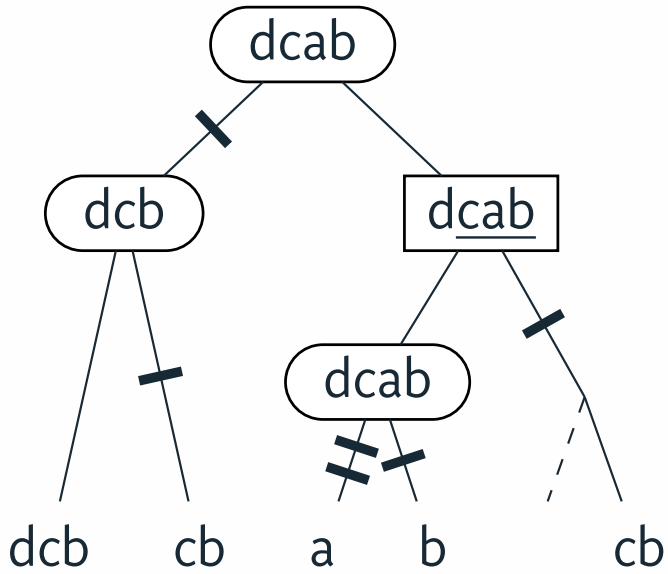
ELEMENTS OF A SUPER-RECONCILIATION



GENERAL SUPER-RECONCILIATION PROBLEM

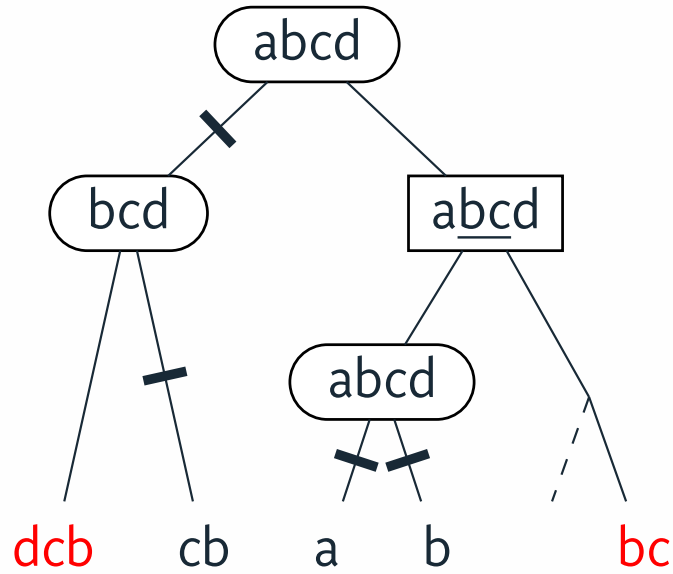
Synteny & species trees ▷ Super-reconciliation minimizing events

ORDERED VS. UNORDERED



ORDERED

- Syntenies are *sequences*
- Relative order of gene families must be *conserved*



UNORDERED

- Syntenies are *sets*
- Any number of *rearrangements* are allowed

PREVIOUS RESULTS[2]

- DL version: Limited set of events
  - Segmental duplications
  - Segmental losses
  - Full losses

- DL-SUPER-RECONCILIATION (ordered)

  $\rightarrow$ *NP-complete*, FPT wrt. t [$t = \#$gene families]

- DL-UNORDERED-SUPER-RECONCILIATION

  $\rightarrow$ *Polynomial:* $O(tn)$ algorithm [$n = \#$nodes in the synteny tree]

[2] Mattéo Delabre et al. "Evolution through segmental duplications and losses: a Super-Reconciliation approach." In: *Algorithms for Molecular Biology* 15.12 (May 2020). DOI: 10.1186/s13015-020-00171-4.

BEYOND DL

Several important evolutionary mechanisms are not modelled in DL:

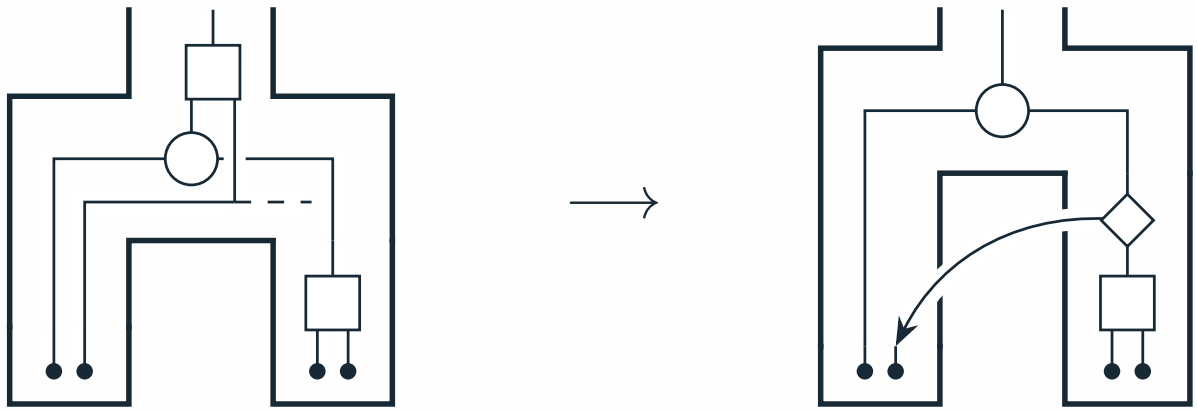
- **Horizontal gene transfers** (HGTs)
- *Tandem gene duplications*
- Gene gains
- Whole genome duplications (WGDs)

OUTLINE

HORIZONTAL GENE TRANSFERS

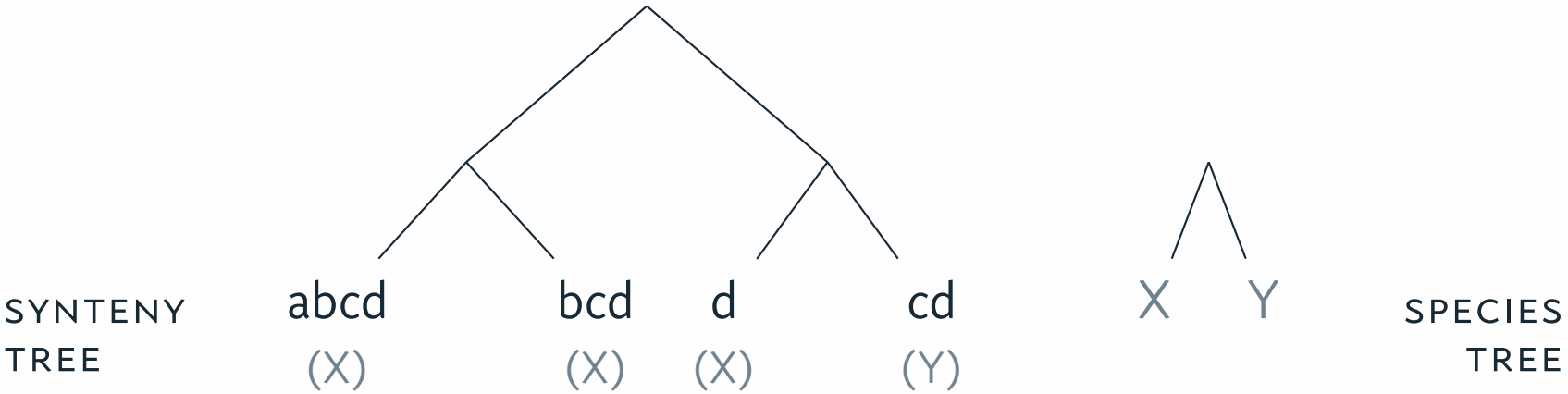
Definition: HORIZONTAL GENE TRANSFER (HGT)

Transfer of genetic material between two *co-existing species*

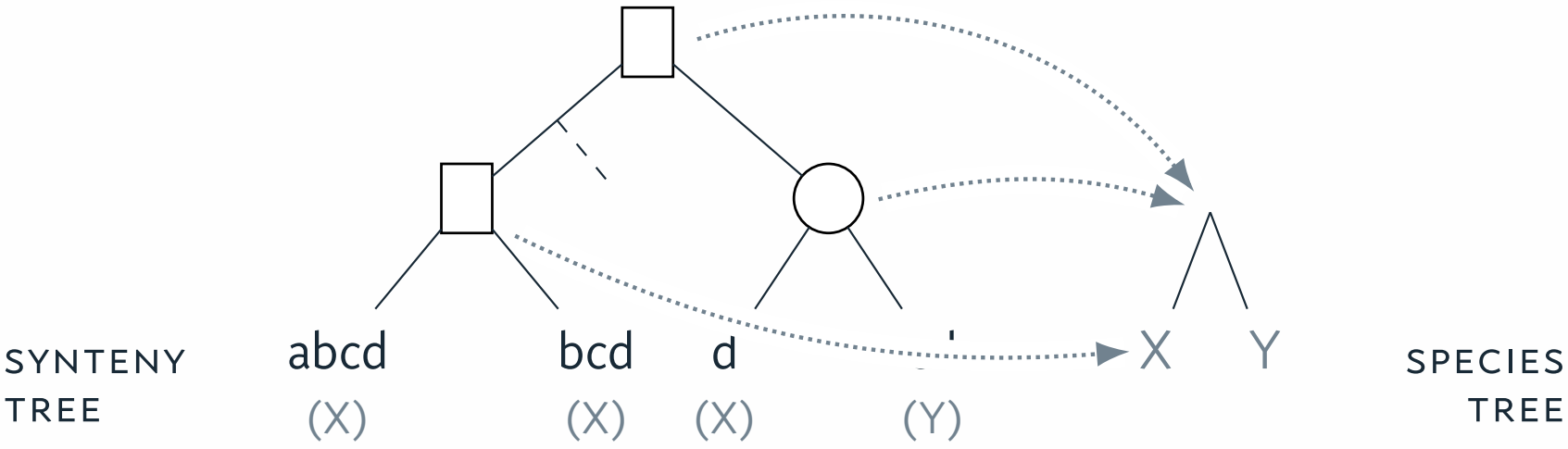


- Similar to segmental duplication, resulting copy in other genome
- *Breaks the coincidence* of tree embeddings in reconciliations

COMBINATORIAL PERSPECTIVE

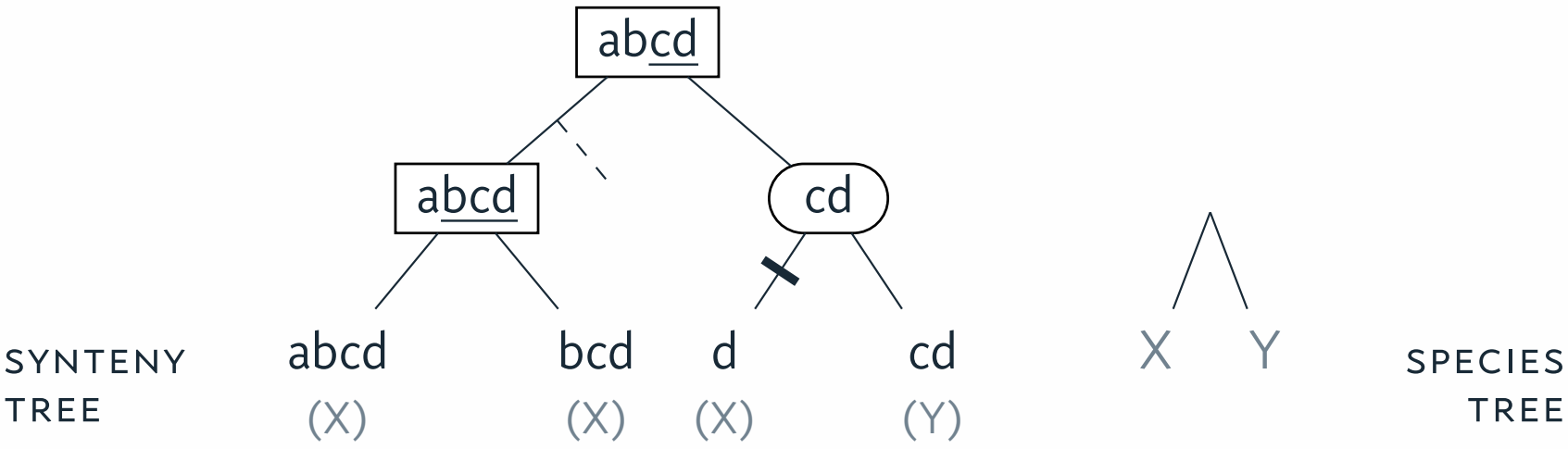


COMBINATORIAL PERSPECTIVE



1. *Synteny tree* → *species tree* mapping
   - Marks each node as a *speciation, duplication, or HGT*
   - Implies *full losses* below some duplications

COMBINATORIAL PERSPECTIVE



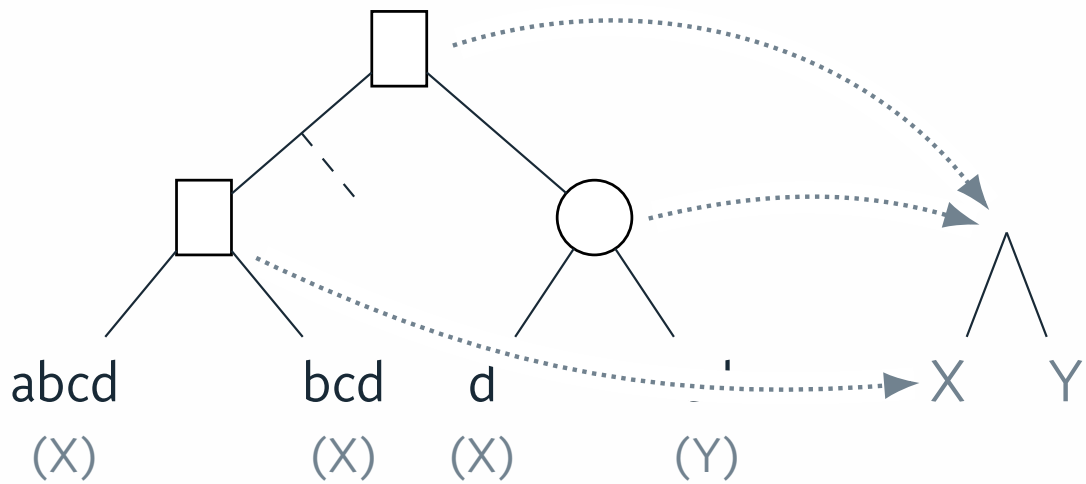
1. *Synteny tree* → *species tree* mapping
   - Marks each node as a *speciation, duplication, or HGT*
   - Implies *full losses* below some duplications

2. Labelling of the synteny tree with *ancestral syntenies*
   - Implies *segmental losses* between some nodes

SYNTENY → SPECIES MAPPING (1/2)

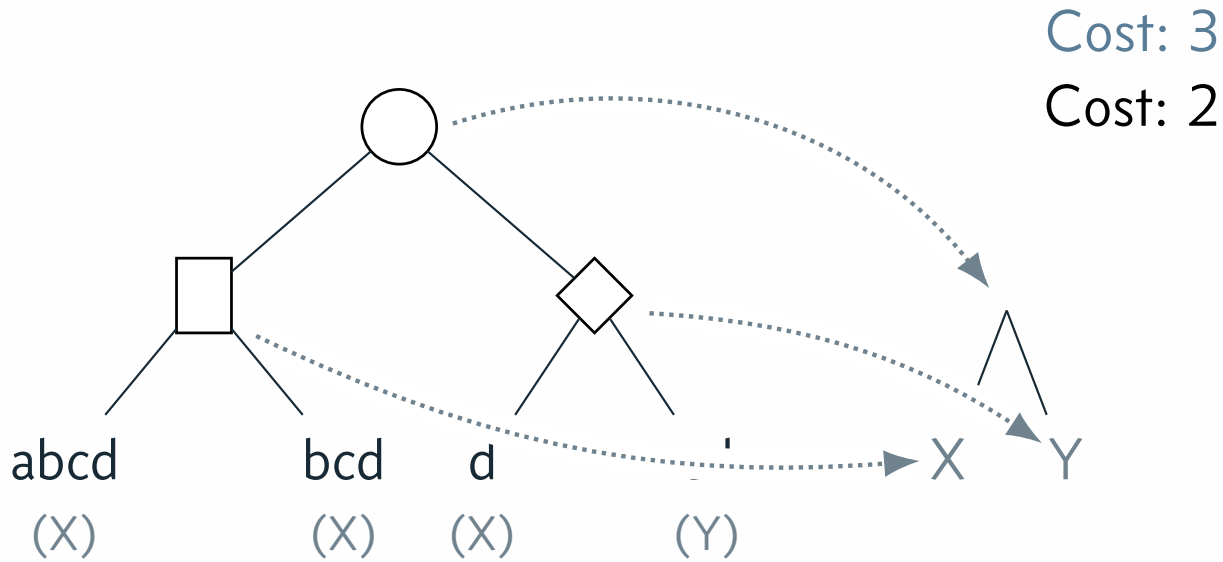
- DL: Only optimal solution is LCA-mapping.
- DTL: LCA-mapping is not always optimal.

Cost: 3



SYNTENY → SPECIES MAPPING (1/2)

- DL: Only optimal solution is LCA-mapping.
- DTL: LCA-mapping is not always optimal.



SYNTENY → SPECIES MAPPING (2/2)

- *Bansal–Alm–Kellis algorithm*[3]*:* Explore all possible mappings (with HGTs allowed) via dynamic programming
- $c[v, \sigma]$: Minimum cost of a mapping for the synteny subtree below v, such that v is mapped to σ
- Complexity: $O(nm)$

[3] Mukul S. Bansal, Eric J. Alm, and Manolis Kellis. "Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss." In: *Bioinformatics* 28.12 (2012), pp. i283–i291. DOI: 10.1093/bioinformatics/bts225.

ANCESTRAL SYNTENY LABELLING

- Label each internal node with an ancestral synteny
  - Minimize the number of lost segments
  - Lost gene families cannot be recovered

ANCESTRAL SYNTENY LABELLING

- Label each internal node with an ancestral synteny
  - Minimize the number of lost segments
  - Lost gene families cannot be recovered

- ORDERED
  - Can be *any subsequence* of a valid ordering of gene families.
  - $O(2^t \cdot t!) = O(2^{t \log t + t})$ options

ANCESTRAL SYNTENY LABELLING

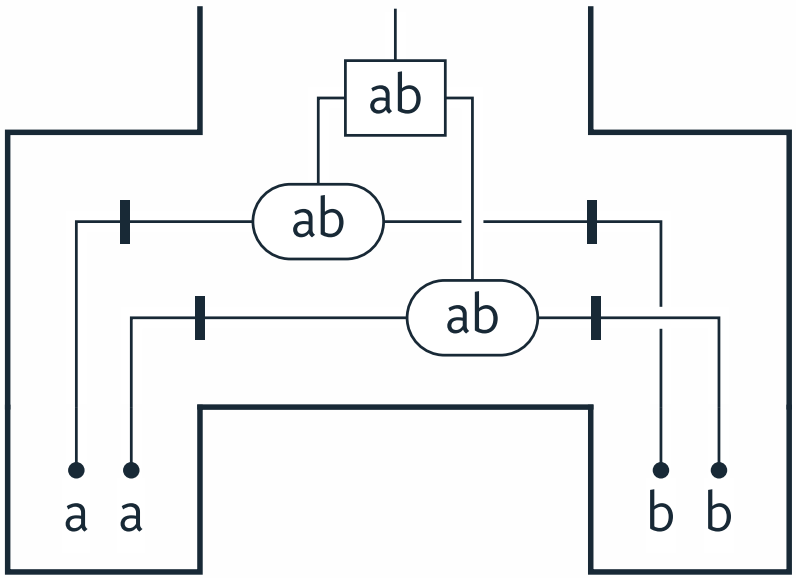
- Label each internal node with an ancestral synteny
  - Minimize the number of lost segments
  - Lost gene families cannot be recovered

- ORDERED
  - Can be *any subsequence* of a valid ordering of gene families.
  - $O(2^t \cdot t!) = O(2^{t \log t + t})$ options

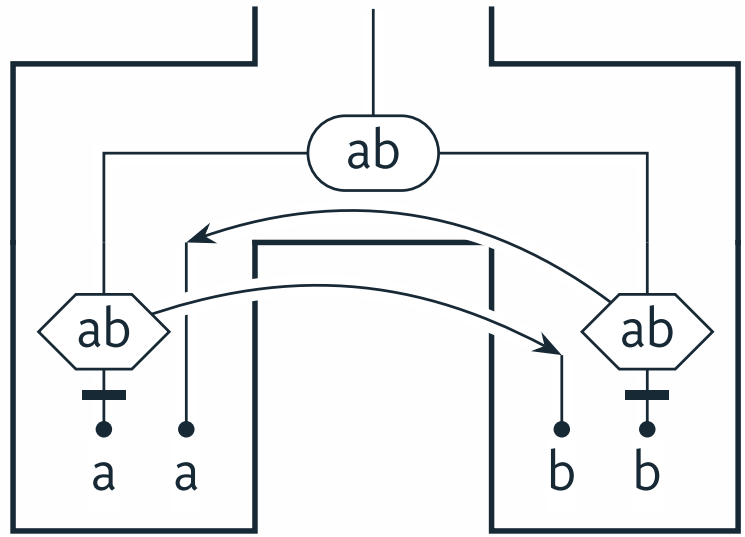
- UNORDERED
  - At most one loss between any two syntenies $\Rightarrow$ 2 options
    1. *Minimal set* of gene families of the subtree
    2. Any *parent synteny set*

TWO-STEPS EXPLORATION

- DL: Events can be minimized in two steps.
- DTL: Two-steps is suboptimal.



#events: 1 + 4 = 5



#events: 2 + 2 = 4

PUTTING IT ALL TOGETHER

- $c[v, \sigma, \chi]$: Minimum cost of a DTL-Super-Reconciliation for the synteny subtree below v, where
  - v is mapped to the σ species
  - v is labelled with the χ synteny

- *Solution:* $\min_{\sigma,\chi} c[\text{root}, \sigma, \chi]$

n: #nodes in synteny tree
m: #nodes in species tree
t: #gene families

- *Complexity:*

ORDERED: $O(nm2^{t \log t + t}) \times O(mt2^t) = O(nm^2 t 2^{t \log t + 2t})$

UNORDERED: $O(nm) \times O(mt) = O(nm^2 t)$

COMPLEXITY SUMMARY

n: #nodes in synteny tree
m: #nodes in species tree
t: #gene families

PROBLEM		COMPLEXITY UP. BOUND
DL	ordered	$O(nt2^{t \log t + 2t})$
	unordered	$O(nt)$
DTL	ordered	$O(nm^2t2^{t \log t + 2t})$
	unordered	$O(nm^2t)$

- Ordered DTL likely NP-complete but still FPT
- Unordered DTL still polynomial

OUTLINE

TANDEM DUPLICATIONS

Definition: TANDEM DUPLICATION (TD)

Local duplication of one or several genes

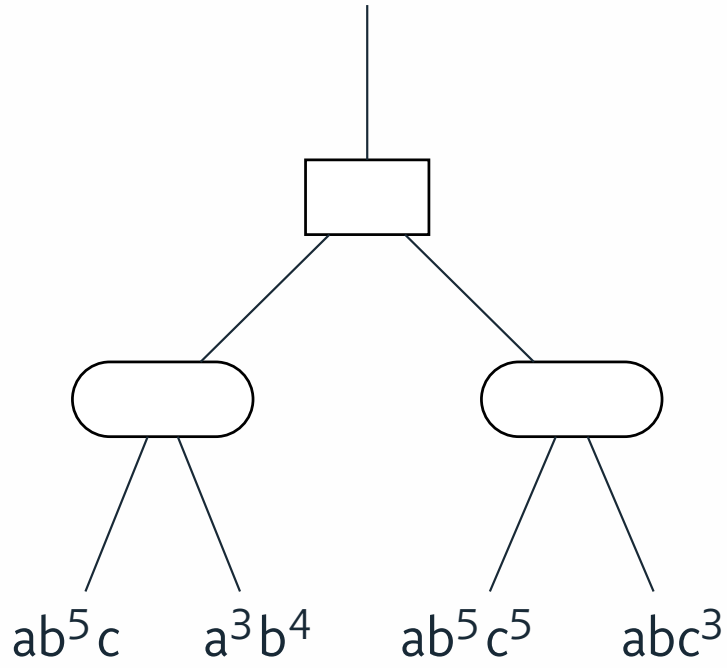
- Allow multiple gene copies in the same synteny: *multisets*
- TDs amplify existing gene content:

$$abc \longrightarrow a^2b^2c$$
$$ab \longrightarrow a^2b^2c \quad \textit{invalid}$$

- There may be successive TDs:

$$a^2bc \xrightarrow{2\ \text{TDs}} a^5bc^2$$

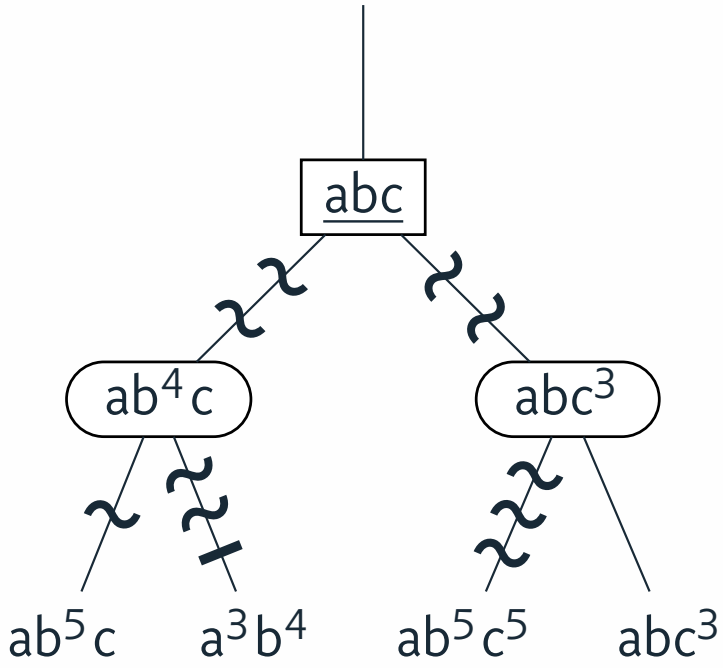
EXAMPLE



− Segmental loss ~ Tandem duplication

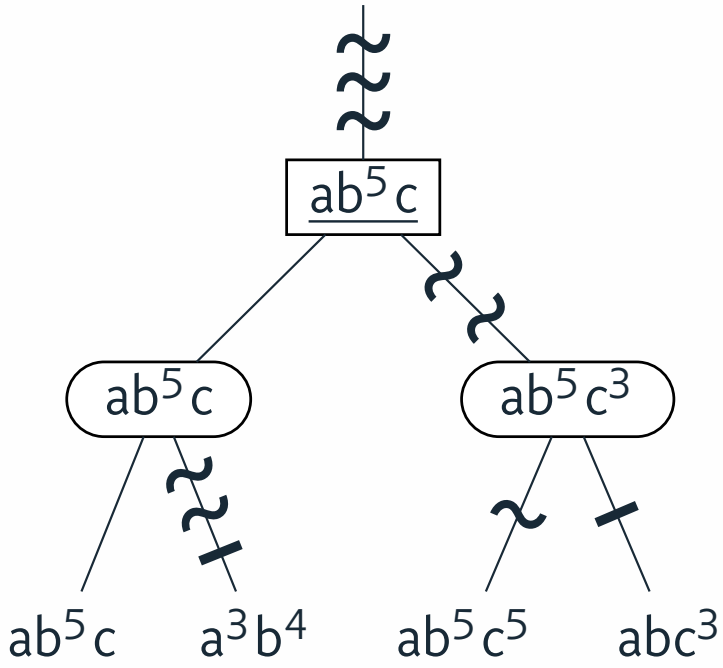
EXAMPLE

Cost: 11



− Segmental loss ~ Tandem duplication

EXAMPLE



Cost: 11

Cost: 10

− Segmental loss ~ Tandem duplication

OUTLINE

KEY POINTS

- Super-Reconciliation *lends itself well to handling HGTs*
  - Requires a more thorough search for optimal solutions
  - Adds a m^2 factor in optimization algorithms
- Further extensions required for applicability
  - *Tandem duplications* (ongoing work)
  - *Gene gains*
- Future work: *Validation* via simulation; *Applications*