Best Matches in large-scale orthology detection

35th TBI Winterseminar in Bled

David Schaller

Max Planck Institute for Mathematics in the Sciences
Bioinformatics Group, University of Leipzig
sdavid@bioinf.uni-leipzig.de

February 11, 2020

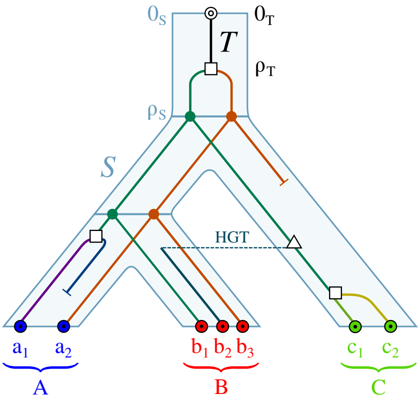
Max Planck Institute for Mathematics in the Sciences

UNIVERSITÄT LEIPZIG

Background

Types of homology:

- **orthologs**
  separated by a speciation event (●),
  similar functions (‘ortholog conjecture’)
- **paralogs**
  separated by a duplication event (□)
- ***(lca-)*xenologs**
  separated by an HGT event (△)



Gene tree *T* (with losses) embedded into the **species tree** *S*.

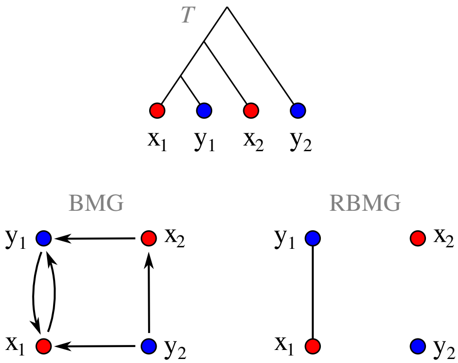
Background – Best Matches

Definition (Best Match[1])

Consider a gene tree T with leaf set $L(T)$ and a surjective color map $\sigma : L(T) \to L(S)$.

Then $y \in L(T)$ is a ***best match*** of $x \in L(T)$ iff $\operatorname{lca}(x, y) \preceq \operatorname{lca}(x, y')$ holds for all leaves y' from species $\sigma(y') = \sigma(y)$.
We write $x \to y$.

If both $x \to y$ and $y \to x$, x and y are ***reciprocal best matches***.



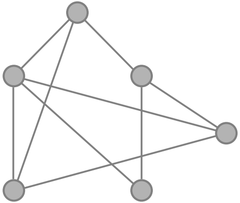
[1] Geiß et al. Best match graphs. *Journal of Mathematical Biology*, 78(7):2015–2057, June 2019.

Background – Best Matches and orthology

Orthology graph Θ ... $xy \in E(\Theta) \iff \mathrm{lca}(x,y)$ was a speciation

- the orthology graph is a **subgraph** of the **RBMG** (if there is no HGT)
  - → no false-negatives

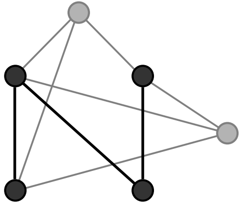
- the orthology graph is a **cograph** $\iff$ P_4-free
  - → the gene tree can be interpreted as a corresponding cotree (speciation = join vertex)
  - → useful for RBMG editing



Background – Best Matches and orthology

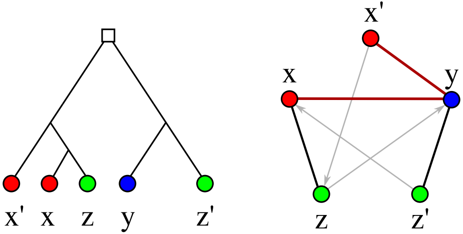
Orthology graph Θ ... $xy \in E(\Theta) \iff \mathrm{lca}(x,y)$ was a speciation

- the orthology graph is a **subgraph** of the **RBMG** (if there is no HGT)
  - → no false-negatives
- the orthology graph is a **cograph** $\iff$ P_4-free
  - → the gene tree can be interpreted as a corresponding cotree (speciation = join vertex)
  - → useful for RBMG editing



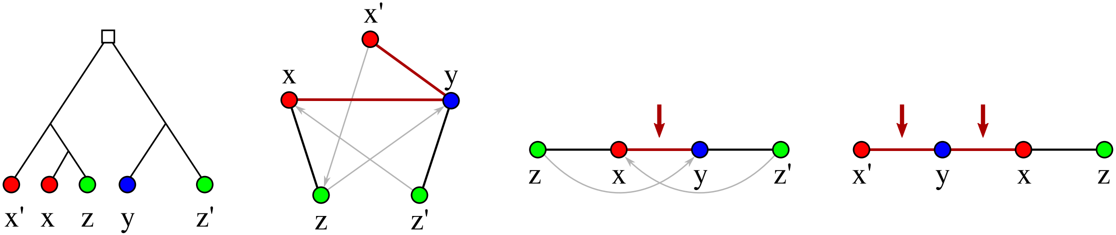
P_4-editing

Identification of false-positive edges w.r.t. orthology



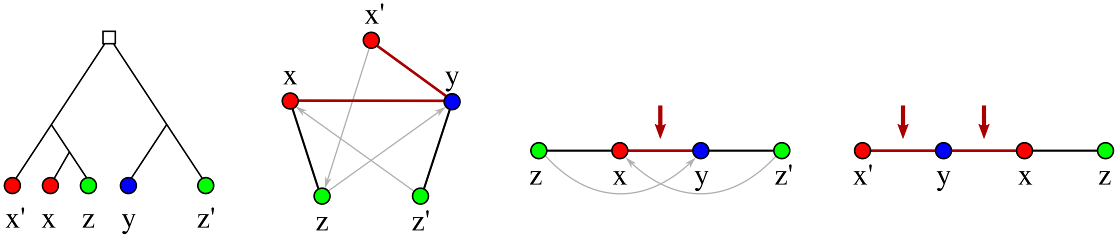
P_4-editing

Identification of false-positive edges w.r.t. orthology



P_4-editing

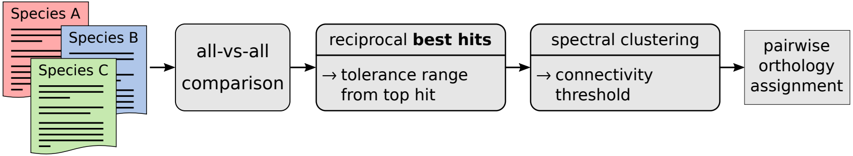
Identification of false-positive edges w.r.t. orthology



- whenever there is a ‘witness species’, we have *good* or *ugly quartets*

Can we use best matches for large-scale orthology detection?

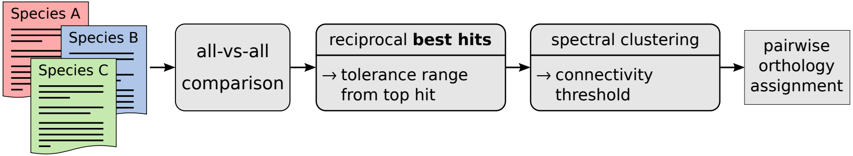
Orthology inference with ProteinOrtho[2]



- **very fast** (all-vs-all comparison with `diamond`, ...)
- best hits $\neq$ best matches, but heuristic via **sub-optimal hits**
- spectral clustering not based on P_4s

[2] Lechner et al. Proteinortho: Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics*, 12(1), 2011.

Orthology inference with ProteinOrtho[2]



Species A

Species B

Species C

all-vs-all comparison

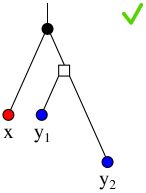
reciprocal **best hits**

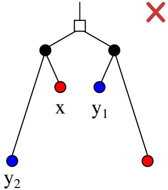
→ tolerance range from top hit

spectral clustering

→ connectivity threshold

pairwise orthology assignment

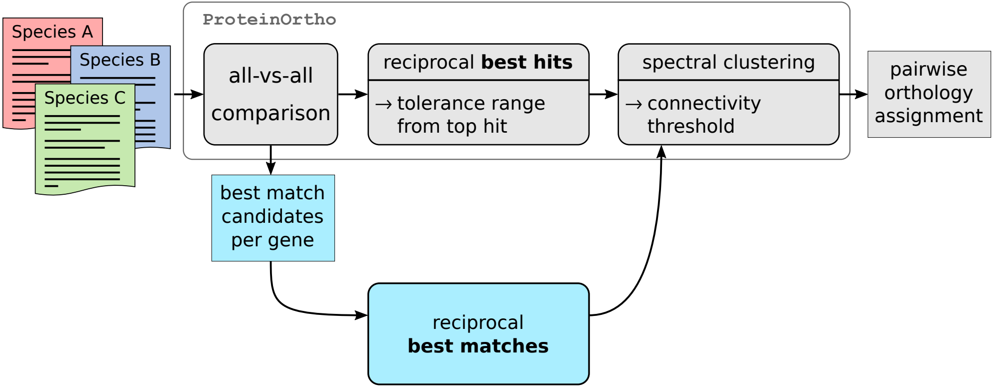




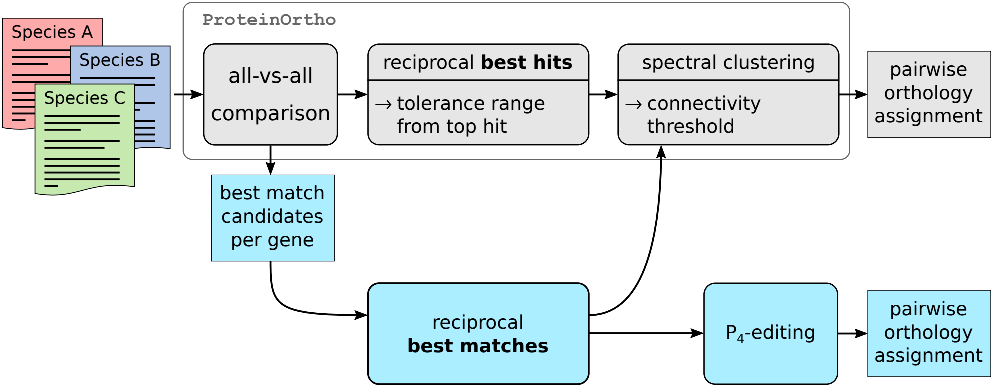
- **very fast** (all-vs-all comparison with `diamond`, ...)
- best hits $\neq$ best matches, but heuristic via **sub-optimal hits**
- spectral clustering not based on P_4s

[2] Lechner et al. Proteinortho: Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics*, 12(1), 2011.

Orthology inference using Best Matches

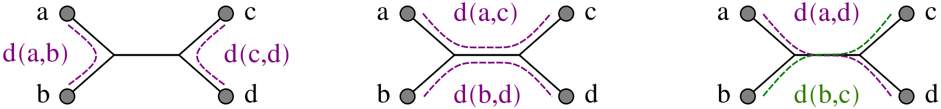


Orthology inference using Best Matches


Species A
Species B
Species C
ProteinOrtho
all-vs-all comparison
reciprocal **best hits**
→ tolerance range from top hit
spectral clustering
→ connectivity threshold
pairwise orthology assignment
best match candidates per gene
reciprocal **best matches**
P_4-editing
pairwise orthology assignment

Best Match inference with quartets

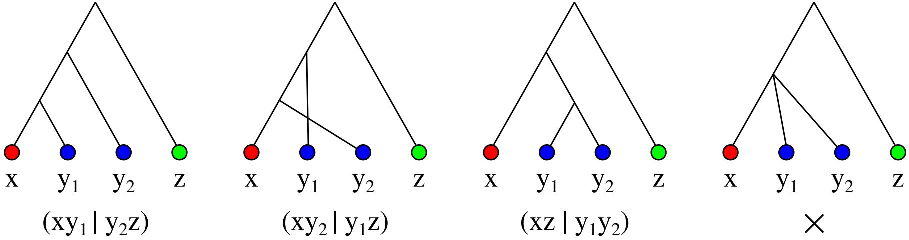
- **quartet relations** can be determined from distance data (using distance sums)



$$d(a,b) + d(c,d) \quad < \quad d(a,c) + d(b,d) \quad \approx \quad d(a,d) + d(b,c)$$

Best Match inference with quartets

- **quartet relations** can be determined from distance data (using distance sums)
- given we have a **trusted outgroup** *z*, there are exactly four trees:



Possible trees on four leaves with a known **outgroup** *z*.

- evaluate all candidate pairs if there are more than two candidates

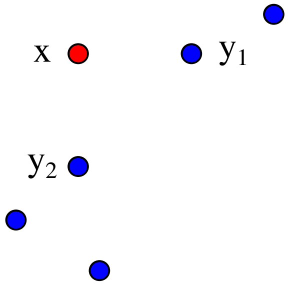
Best Match inference with quartets

To infer the best matches of gene *x* we need:

Best Match inference with quartets

To infer the best matches of gene *x* we need:

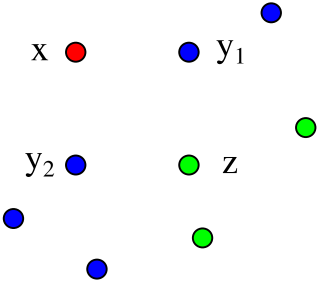
- candidate genes $(y_1, y_2, \ldots)$
  - → blast hits above a certain E-value



Best Match inference with quartets

To infer the best matches of gene x we need:

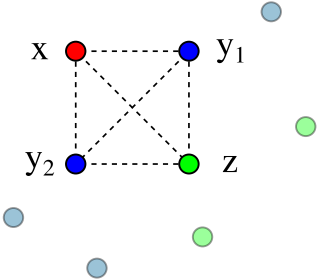
- candidate genes ($y_1, y_2, \ldots$)
  - → blast hits above a certain E-value
- outgroup genes ($z, \ldots$)
  - → blast hits from outgroup species as heuristic
  - → species tree required



Best Match inference with quartets

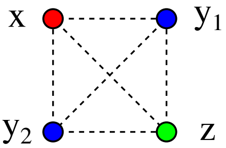
To infer the best matches of gene x we need:

- candidate genes $(y_1, y_2, \ldots)$
  - → blast hits above a certain E-value
- outgroup genes $(z, \ldots)$
  - → blast hits from outgroup species as heuristic
  - → species tree required
- distances



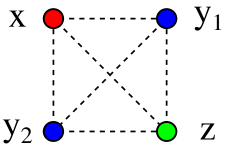
Getting distances

- **Idea I: Realignment**
  - → exact local or global alignments of all required sequence pairs
  - → given a sequence evolution model / rate matrix:
    compute maximum likelihood distance
  - → possible, but a bottleneck



Getting distances

- **Idea I: Realignment**
  - → exact local or global alignments of all required sequence pairs
  - → given a sequence evolution model / rate matrix: compute maximum likelihood distance
  - → possible, but a bottleneck
- **Idea II: Bitscores**
  - → infer quartet topology from bitscores
  - → transformation into distances?
  - → length normalization, missing values, ...?



Species tree

- **Case I: rooted species tree available (from database, ...)**
  → great!

- **Case II: rooted species tree not available**
  → inference from orthology / paralogy relations: `ParaPhylo`[3]
  → e.g. based on `ProteinOrtho` ouput
  → limited to data sets of approx. 20 species
  → replace ILP steps by heuristics

[3] Hellmuth et al. Phylogenomics with Paralogs. *PNAS*, 112(7):2058–2063, 2015.

Species tree

- **Case I: rooted species tree available (from database, ...)**

  → great!

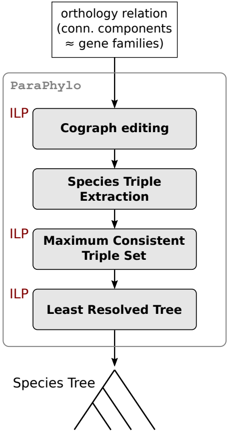
- **Case II: rooted species tree not available**

  → inference from orthology / paralogy relations: `ParaPhylo`[3]

  → e.g. based on `ProteinOrtho` ouput

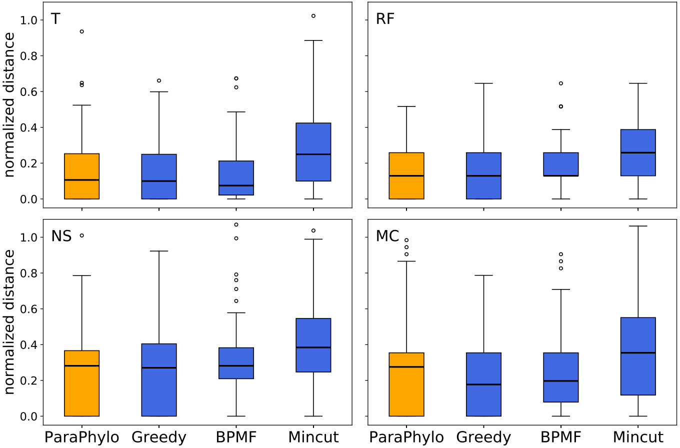
  → limited to data sets of approx. 20 species

  → replace ILP steps by heuristics


orthology relation (conn. components ≈ gene families)
ParaPhylo
ILP
Cograph editing
Species Triple Extraction
ILP
Maximum Consistent Triple Set
ILP
Least Resolved Tree
Species Tree

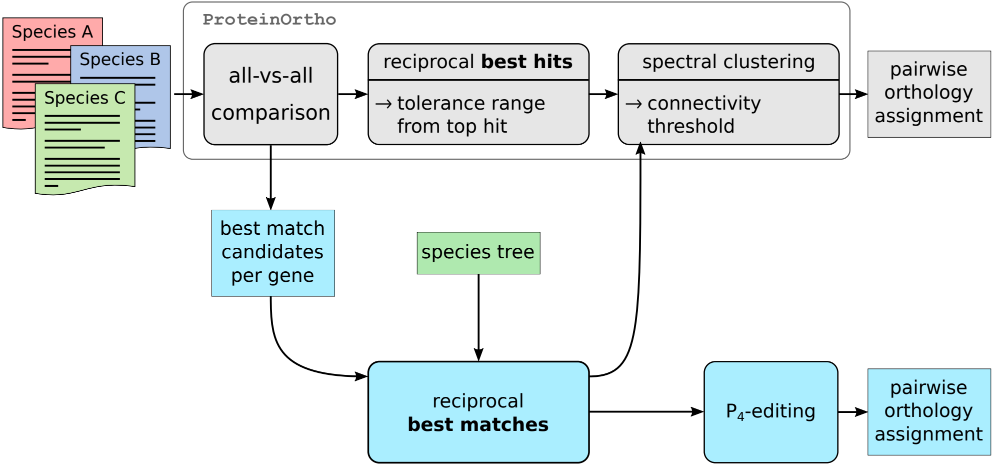
[3] Hellmuth et al. Phylogenomics with Paralogs. *PNAS*, 112(7):2058–2063, 2015.

Species tree: some results



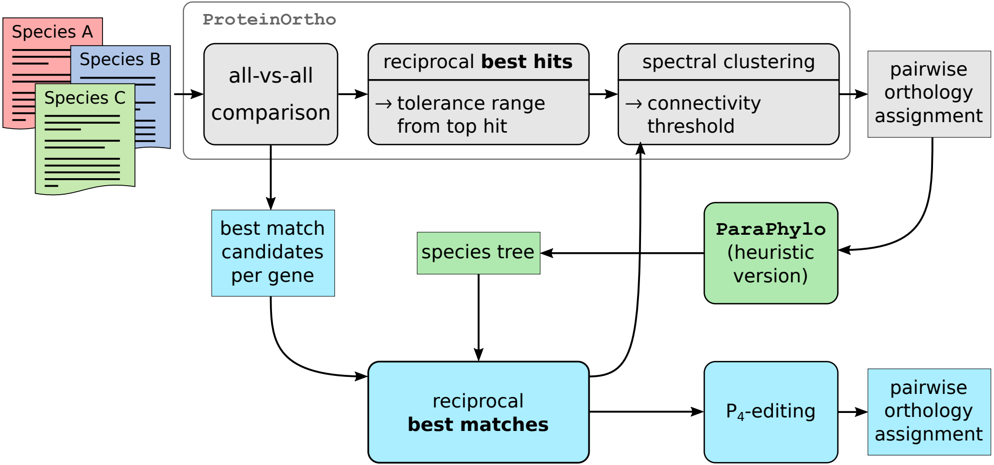
- 100 scenarios à 10 species
- simulated sequences for 1000 gene families
- orthology estimation with `ProteinOrtho`
- various tree distance metrics
  (Triple metric, Robinson-Foulds, Nodal Splitted, Matching Cluster)

Summary

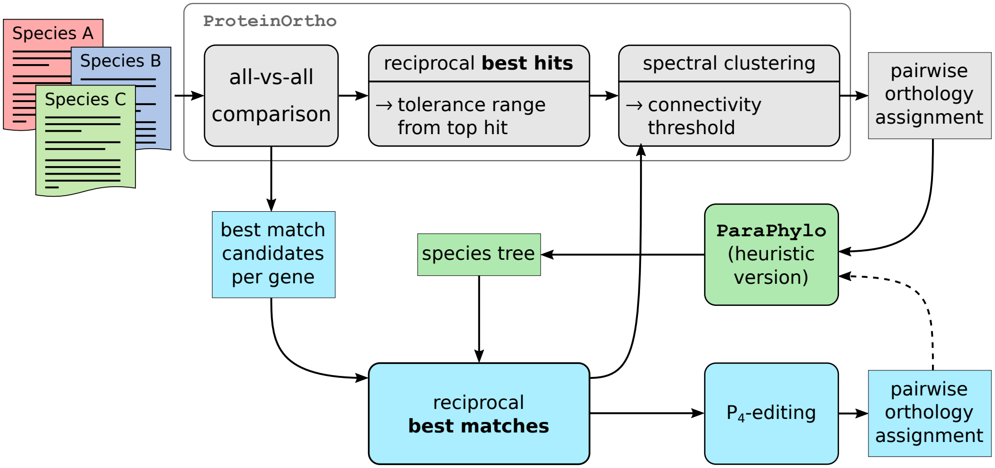


Species A
Species B
Species C
ProteinOrtho
all-vs-all comparison
reciprocal **best hits**
→ tolerance range from top hit
spectral clustering
→ connectivity threshold
pairwise orthology assignment
best match candidates per gene
species tree
reciprocal **best matches**
P_4-editing
pairwise orthology assignment

Summary

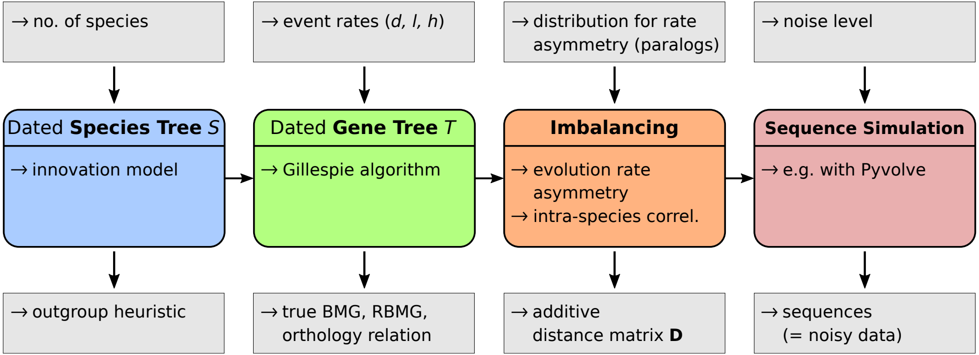

Species A
Species B
Species C
ProteinOrtho
all-vs-all comparison
reciprocal **best hits**
→ tolerance range from top hit
spectral clustering
→ connectivity threshold
pairwise orthology assignment
best match candidates per gene
species tree
ParaPhylo (heuristic version)
reciprocal **best matches**
P_4-editing
pairwise orthology assignment

Summary


Species A
Species B
Species C
ProteinOrtho
all-vs-all comparison
reciprocal **best hits**
→ tolerance range from top hit
spectral clustering
→ connectivity threshold
pairwise orthology assignment
best match candidates per gene
species tree
ParaPhylo (heuristic version)
reciprocal **best matches**
P_4-editing
pairwise orthology assignment

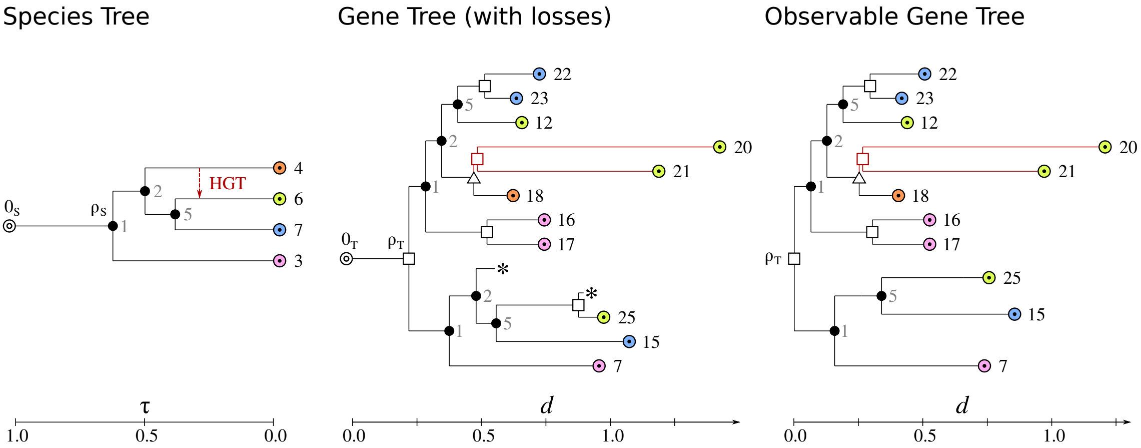
Thank you for your attention!

Appendix: Simulation of Distance Data



→ no. of species
→ event rates (d, l, h)
→ distribution for rate asymmetry (paralogs)
→ noise level
Dated **Species Tree** S
→ innovation model
Dated **Gene Tree** T
→ Gillespie algorithm
Imbalancing
→ evolution rate asymmetry
→ intra-species correl.
Sequence Simulation
→ e.g. with Pyvolve
→ outgroup heuristic
→ true BMG, RBMG, orthology relation
→ additive distance matrix **D**
→ sequences (= noisy data)

Appendix: Simulation of Distance Data



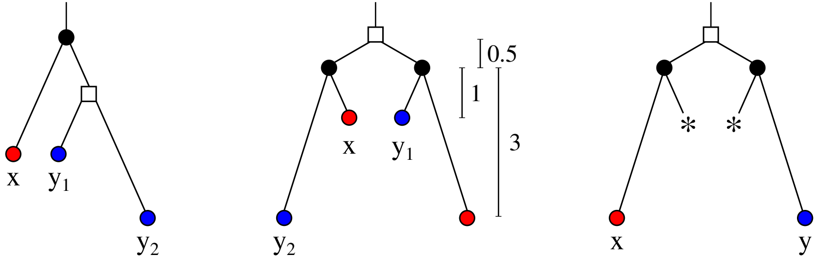
Species Tree

Gene Tree (with losses)

Observable Gene Tree

Appendix: Best Matches vs Best Hits

- y is a **best hit** of x if $d(x, y) \leq d(x, y')$ holds for all leaves y' from species $\sigma(y') = \sigma(y)$
- orthology assessment: **Reciprocal best hits (RBH)** or **reciprocal best matches (RBM)**?

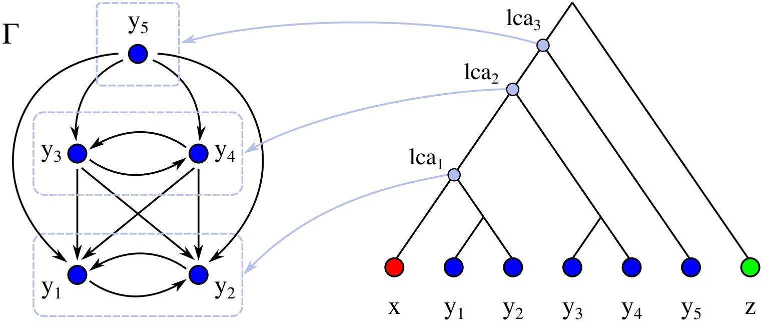


RBH	partially	✗	✗
RBM	✓	✓	sometimes

Appendix: Best Match inference with quartets – auxiliary graph

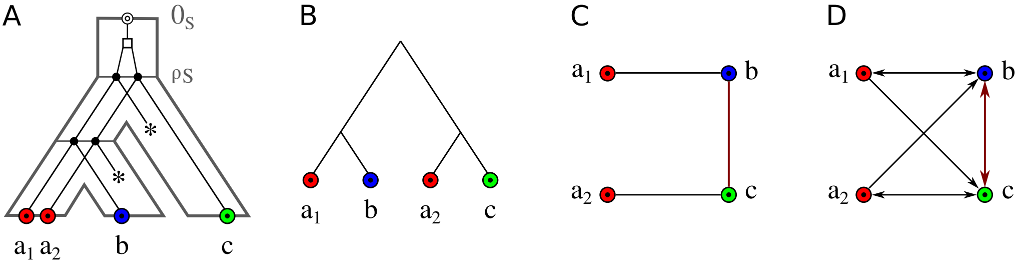
- consider all pairs and construct a digraph Γ on the set of candidates Y of species s

  → $(y'', y') \in E(\Gamma)$ iff $\mathrm{lca}(x, y') \preceq \mathrm{lca}(x, y'')$



Example auxiliary graph Γ for best inference of a gene x in species s (blue).

Appendix: Differential gene loss


A
0_S
ρS
*
*
a_1 a_2
b
c
B
a_1
b
a_2
c
C
a_1
b
a_2
c
D
a_1
b
a_2
c