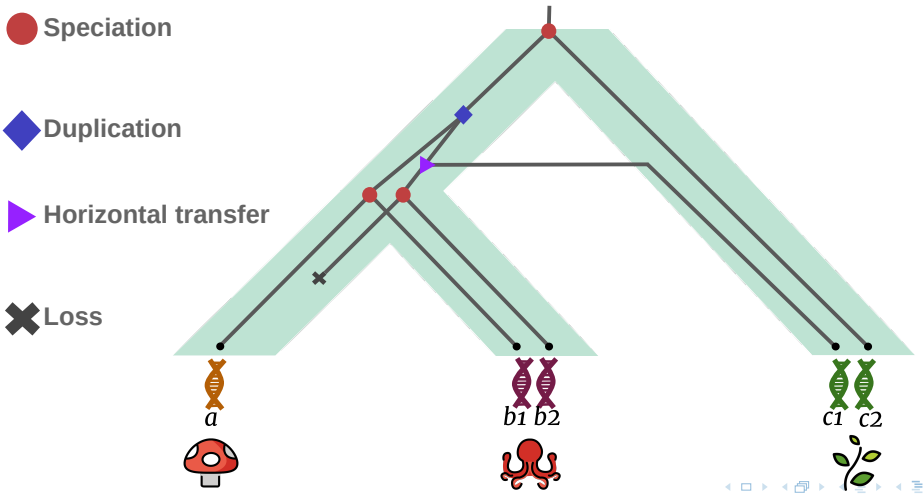

$40^{th}$ TBI Winterseminar in Bled | Feb 11, 2025

# Reconstruction of evolutionary scenarios containing horizontal gene transfer

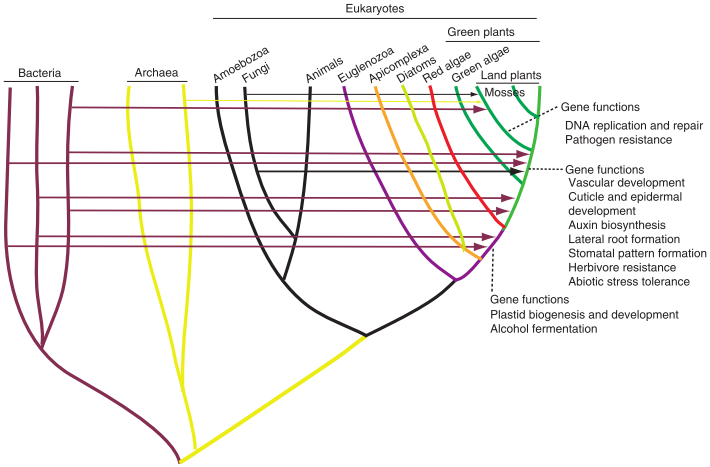
José Antonio Ramírez Rafael

**Joint work:** Annachiara Korchmaros, Maribel Hernández Rosales and Peter Florian Stadler

# Evolutionary scenarios: gene, species, and evolutionary events

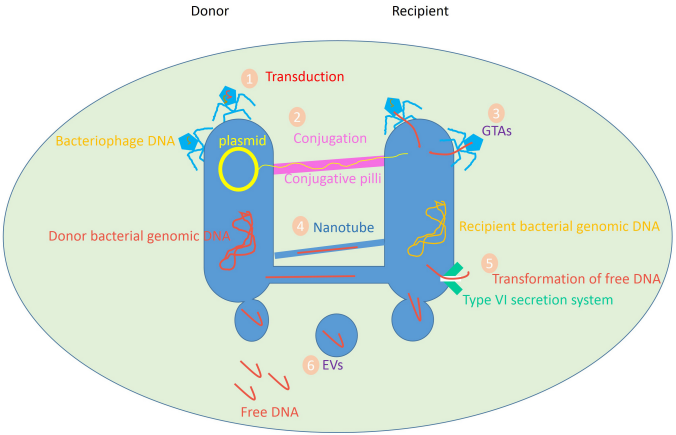


# Horizontal gene transfer: key process in biology



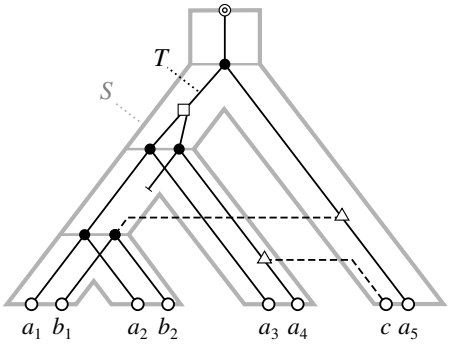

[?]

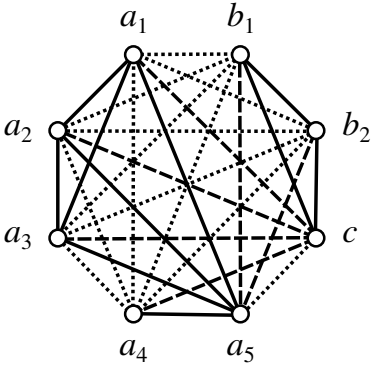
# Horizontal gene transfer: key process in biology



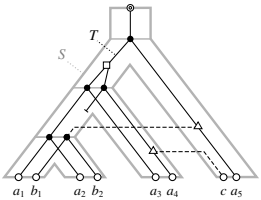

[?]

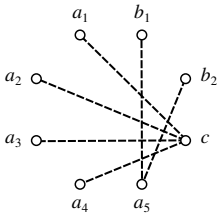
# Graph 3-partition: Measurable gene-to-gene relations



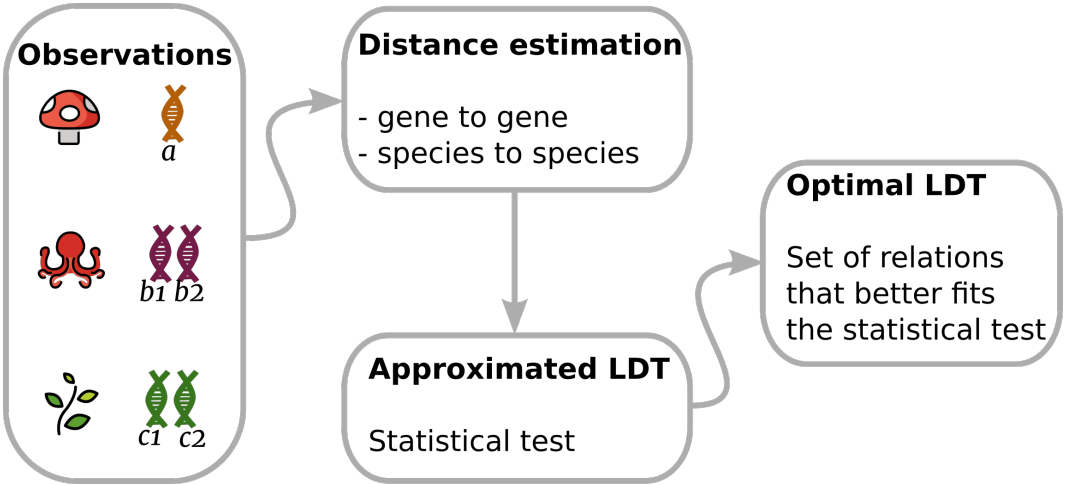
# Graph 3-partition | Later Divergence Time (LDT)




Two genes $a, b$ are in a ***LDT relationship*** iff they diverged after the species $\sigma(a), \sigma(b)$ where they reside.
Every LDT relation is evidence for an HGT event.

# From observations to LDT graphs



## Data: partial estimation of distances

Given a set of homologous genes $V$,
we have a **partial** set of measurements $D \subseteq E = V \times V$ such that

- $d_{uv}$ is a distance between genes $a, b$, and
- $\delta_{uv}$ is a distance between species $\sigma(a), \sigma(b)$.

for $ab \in D$.

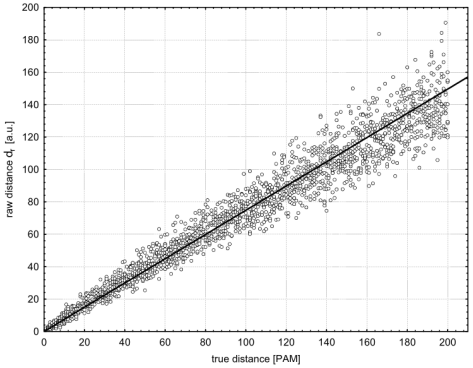
In particular, we use **scoredist**; a **gene-to-gene** distance estimate based on bit-score of alignment hits.

**Species-to-species** distance is estimated as the **mean scoredist** between genes in the corresponding species

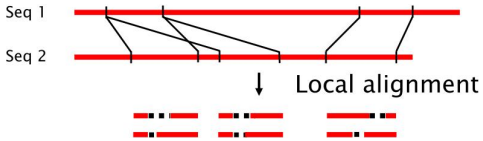
# Scoredist: a simple approximation of number of substitutions

*Scoredist*: $d_r = -\log(z) \cdot 100$

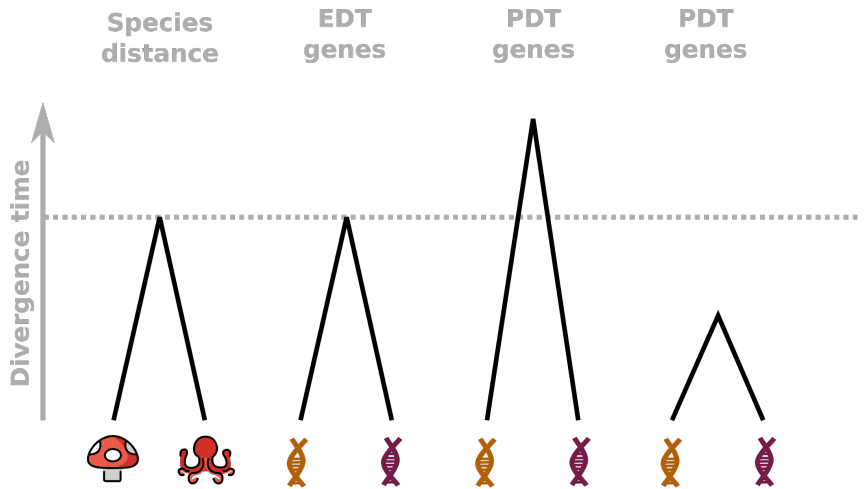
Where $z \in [0, 1]$ is the normalized alignment score between two sequences.



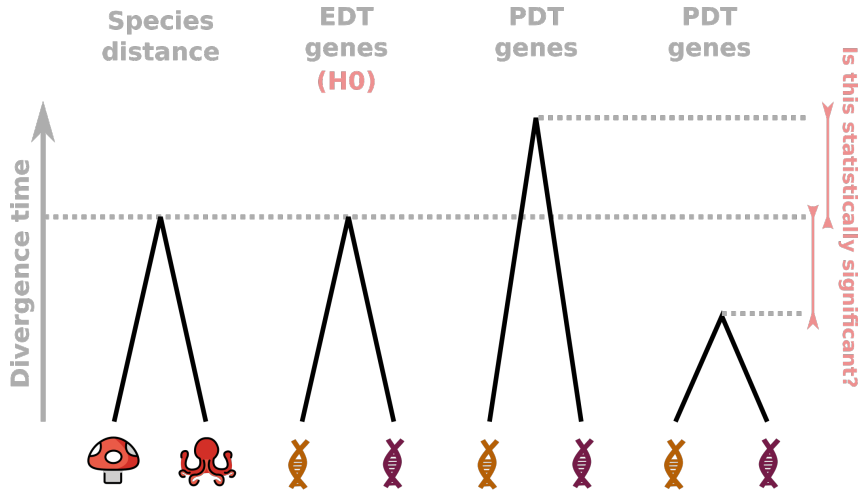

[?]



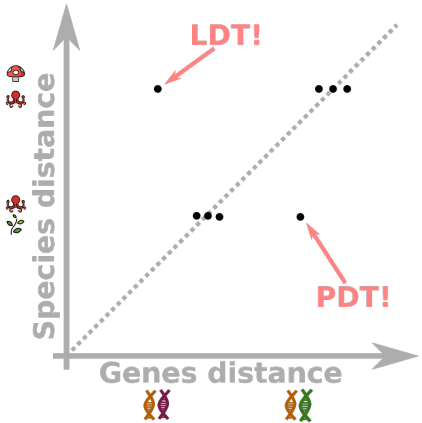
# Estimated distances allow us to approximate a G3P



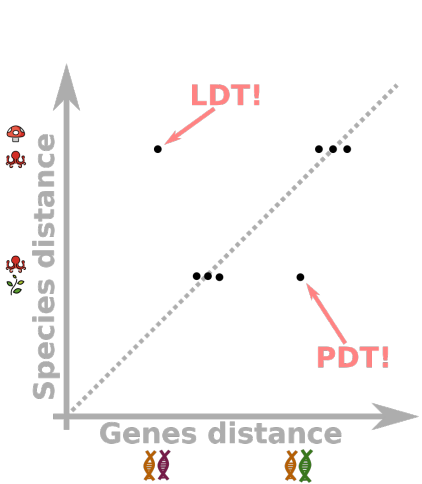
# Estimated distances allow us to approximate a G3P



# Statistical approach: outlier detection



# Statistical approach: outlier detection




**Slope estimation**

Theil–Sen

$$(y_j - y_i)/(x_j - x_i)$$

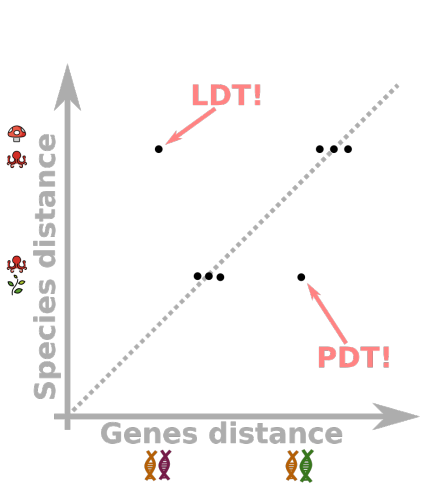
DLIGHT

$$\text{median}(y)/\text{median}(x)$$

MLE

$$L(D, \vec{\beta}) = \sum_i (y_i - \vec{\beta} \cdot \vec{x_i})^2$$

# Statistical approach: outlier detection




## Slope estimation

Theil–Sen

$$(y_j - y_i)/(x_j - x_i)$$

DLIGHT

$$\text{median}(y)/\text{median}(x)$$

MLE

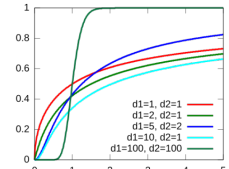
$$L(D, \vec{\beta}) = \sum_i (y_i - \vec{\beta} \cdot \vec{x_i})^2$$

## Outlier detection

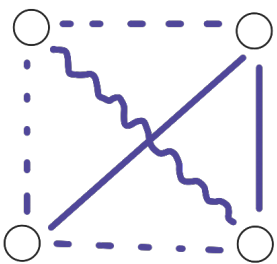
Cook Distance

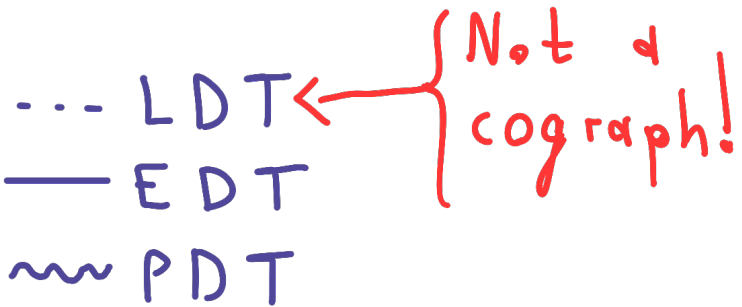
$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{ps^2}$$

p-values determined by F-distribution



## Inferred gene-to-gene relations are not a G3P!





# ILP approach for correcting gene-to-gene relations: objective function

Let $V$ be a set of homolog genes, and $\tilde{E} \subseteq V \times V$ the LDT edges predicted by outliers

Binary variables

- $e_{uv}$ for $uv \in V \times V$
- $T_{ABIC}$, $T'_{ABIC}$ for $A, B, C \in \mathcal{T}(V)$ s.t. $|\{A, B, C\}| = 3$

Binary constants

- $\tilde{e}_{uv} = \begin{cases} 1 & \text{if } uv \in \tilde{E} \\ 0 & \text{otherwise} \end{cases}$

Objective functions (Minimize)

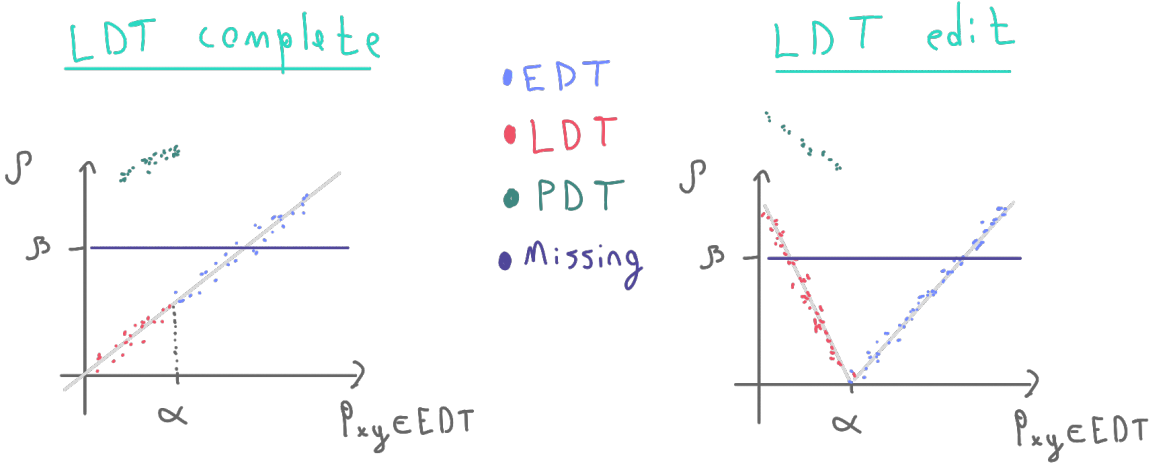
LDT complete

$$\sum_{xy \in E} f_{xy} e_{xy}$$

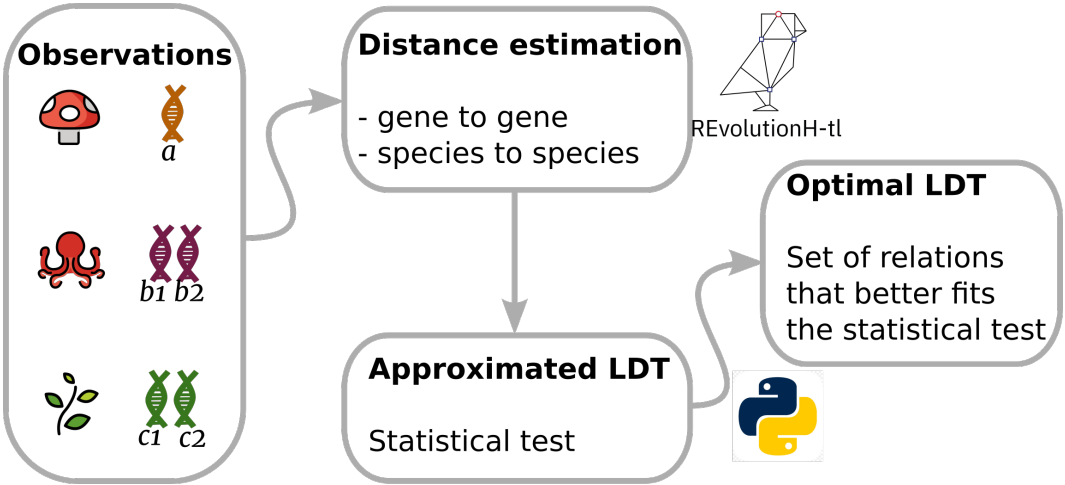
LDT edit

$$\sum_{xy \in E} r_{xy} \left[ (1 - e_{xy}) \tilde{e}_{xy} + (1 - \tilde{e}_{xy}) e_{xy} \right]$$

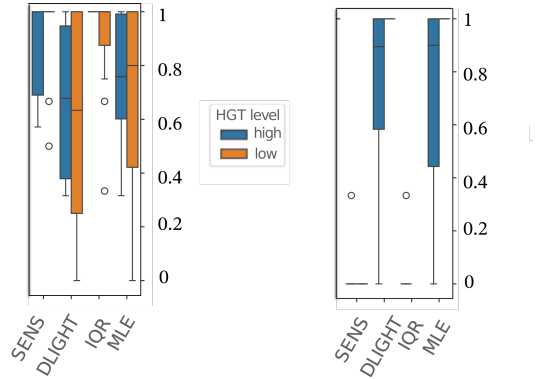
# ILP approach for correcting gene-to-gene relations: costs



# Package for inferring Horizontal gene transfer ($\alpha$-version)



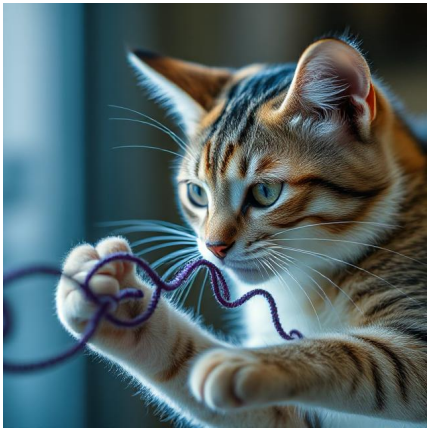
# High precision of predictions



## Take home notes

- Graph theory reveals combinatorial restrictions on evolutionary relationships
- Our methodology can deal with noisy gene distance estimates
- Our methodology can deal with missing data points
- ILP problem relaxation allows us to find near-optimal LDT graphs
- ILP solvers are slow... let's make some heuristics!
- Coming soon: Use inferred relations to infer the direction of the transfer

# Thanks!



AI generated: *Cat saying thanks after a conference about horizontal gene transfer in Bled, Slovenia.*