Space of Gene/Species Trees Reconciliations and Probabilistic Models

Jean-Philippe Doyon[1,2] Cedric Chauve[3] Sylvie Hamel[2]

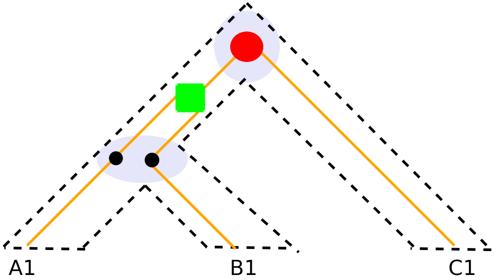
1- LIRMM, Université Montpellier 2 and CNRS
2- Département d'Informatique et de Recherche Opérationnelle,
Université de Montréal
3- Department of Mathematics, Simon Fraser University

Integrative Post-Genomics
Lyon, November 2009

Gene Family Evolution

The evolution of a genome is determined by

Speciation (🔴), Duplication (🟩), and Loss (●).



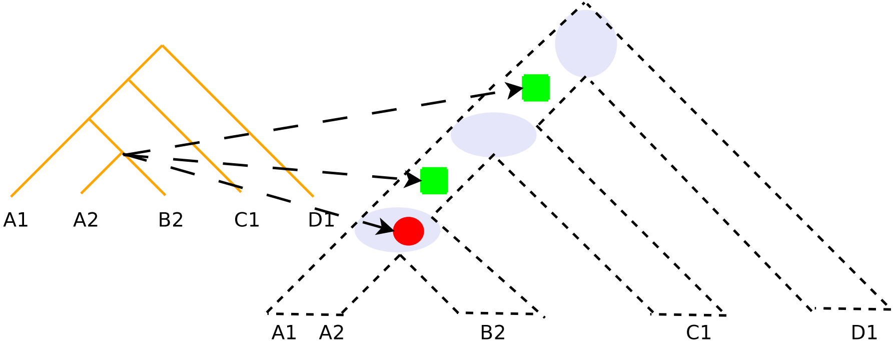
Why it is important to study the evolution of homologous genes?

- Orthologous and paralogous genes
- Gene content of ancestral genomes
- Phylogenomic

The Problem

The Main Question

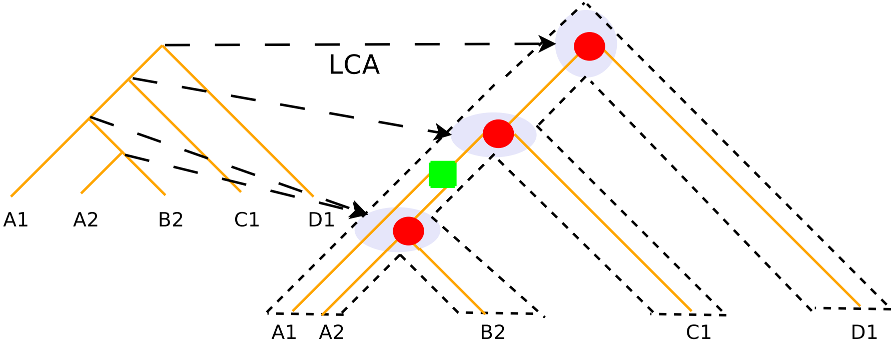
Define the evolution of the **Gene Tree** according to **Species Tree** in term of speciation, duplication, and loss events.



Definitions

The Most Parsimonious Reconciliation (α_{min})

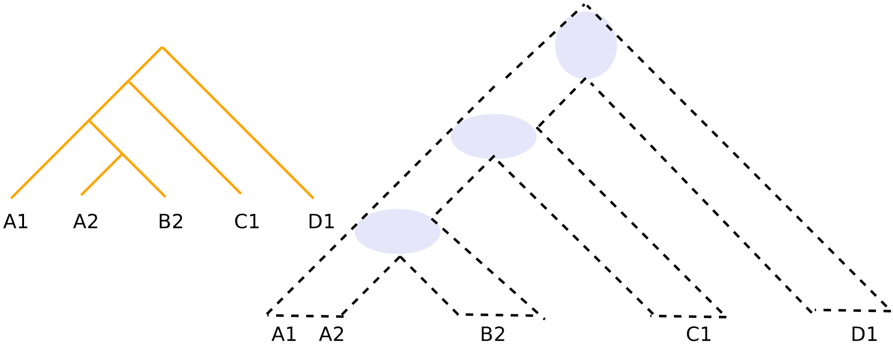
maps a gene *u* (of *G*) the lowest possible in *S*



A more General Definition

A Reconciliation between *G* and *S*

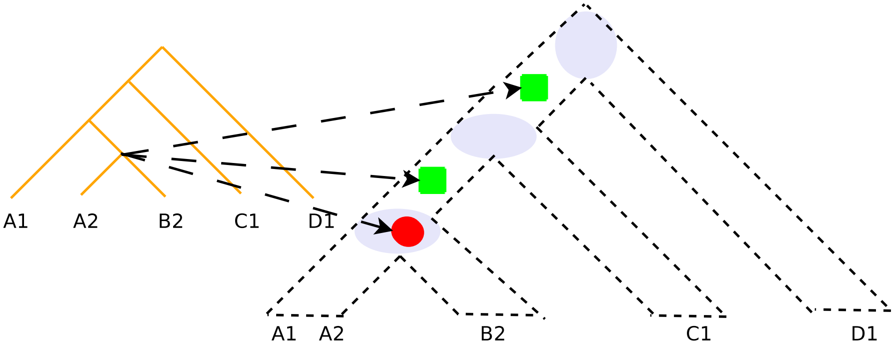
- Each internal node is mapped either on the LCA or on an edge above.
- Descendance Relationships.



A more General Definition

A Reconciliation between *G* and *S*

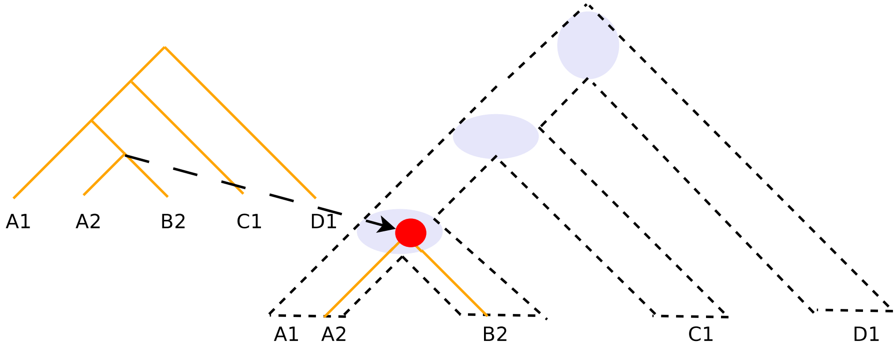
- Each internal node is mapped either on the LCA or on an edge above.
- Descendance Relationships.



A more General Definition

A Reconciliation between *G* and *S*

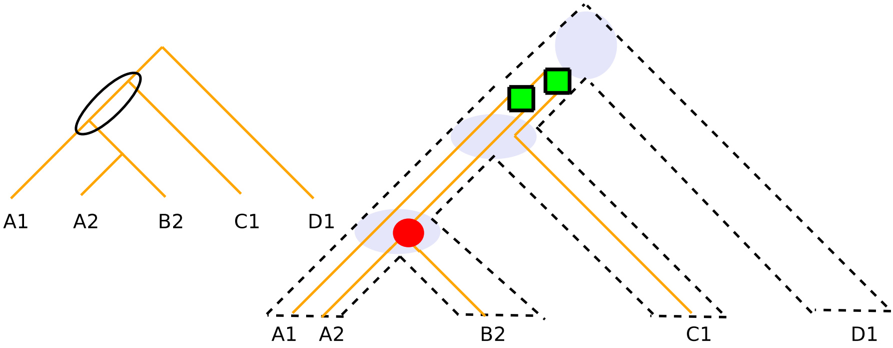
- Each internal node is mapped either on the LCA or on an edge above.
- Descendance Relationships.



A more General Definition

A Reconciliation between *G* and *S*

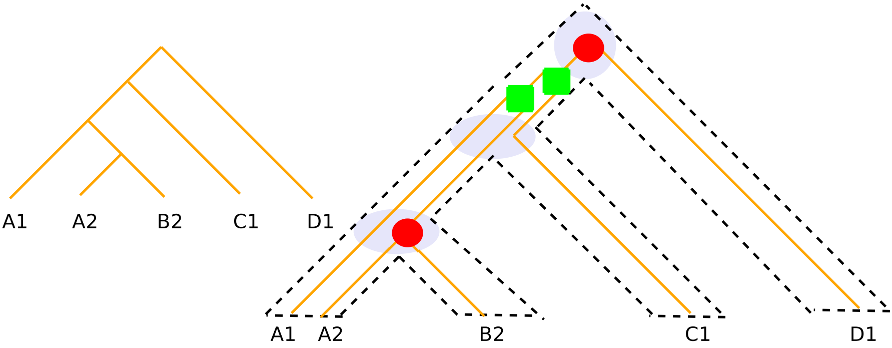
- Each internal node is mapped either on the LCA or on an edge above.
- Descendance Relationships.



A more General Definition

A Reconciliation between *G* and *S*

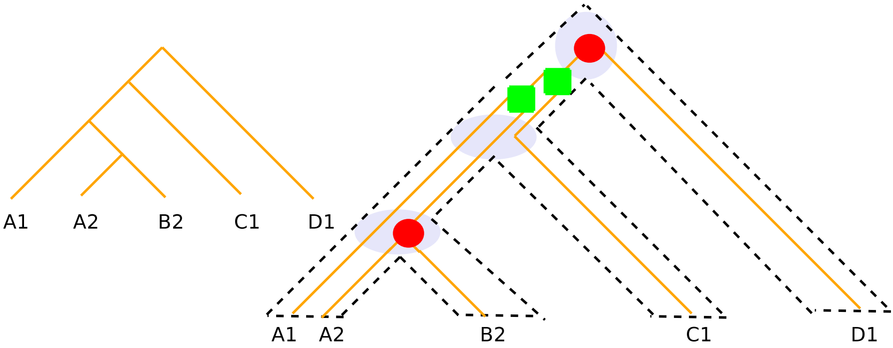
- Each internal node is mapped either on the LCA or on an edge above.
- Descendance Relationships.



A more General Definition

Properties

- Does not ONLY induce the LCA reconciliation.
- The number of reconciliations is finite, but can be exponential.



Reconciliation Space Exploration

This simple definition allows

- Count the number of reconciliations.
- Generate randomly and uniformly a reconciliation.
- Define operators used to explore the whole space.
- Exhaustively explore the space.

Counting and Randomization algorithms

Counting the Number of Reconciliations

- Dynamic programming algorithm in $O(|G||S|)$ time and space.

Randomly Generate a reconciliation

- Algorithm in $O(|G||S|)$ time and space.
- Uniform distribution over all reconciliations.

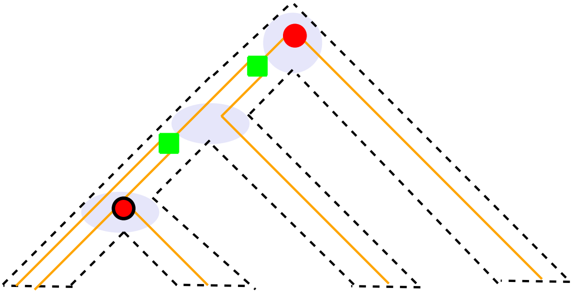
Nearest Mapping Change

Upward NMC

- Changes a speciation into a duplication.
- Moves a duplication upward.

Downward NMC

- Changes a duplication into a speciation.
- Moves a duplication downward.



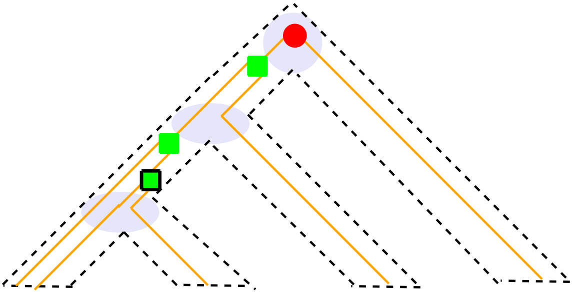
Nearest Mapping Change

Upward NMC

- Changes a speciation into a duplication.
- Moves a duplication upward.

Downward NMC

- Changes a duplication into a speciation.
- Moves a duplication downward.



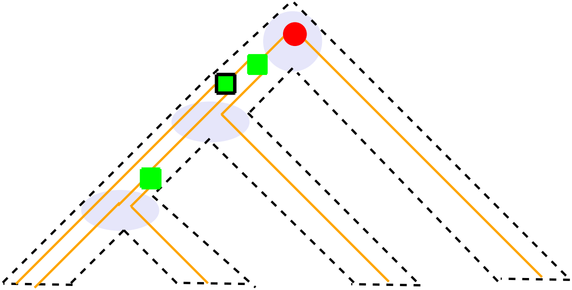
Nearest Mapping Change

Upward NMC

- Changes a speciation into a duplication.
- Moves a duplication upward.

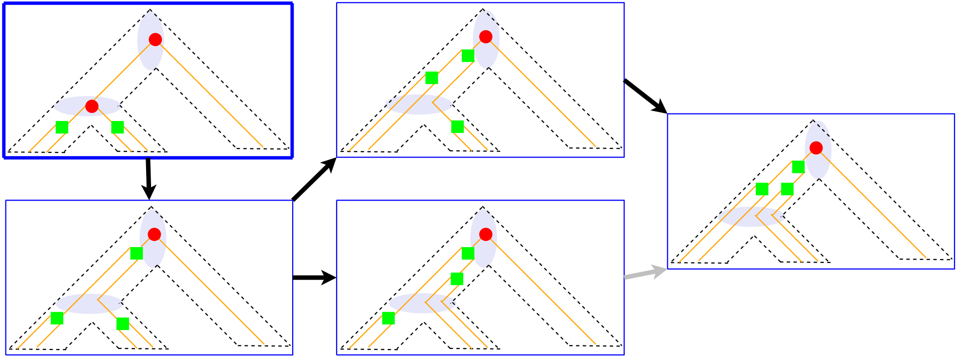
Downward NMC

- Changes a duplication into a speciation.
- Moves a duplication downward.



Sufficient to explore the whole space of reconciliations.

Exhaustive Exploration



- Architecture rooted at α_{min};
- Exploration of the whole space of reconciliations in time $\Theta(\#rec)$.

Probabilistic framework

Input data

- A gene tree G and a species tree S.
- Branch lengths (in time) and duplication/loss rates along S.

Posterior Probability $P(\alpha|G)$

$P(G, \alpha)$ is the probability that the evolution of a gene along S generates

- the Gene Tree G;
- and the Reconciliation α.

$$\begin{aligned} P(\alpha|G) &= \frac{P(G,\alpha)}{P(G)} \\ &= \frac{P(G,\alpha)}{\sum_{\alpha' \in \mathcal{T}} P(G,\alpha')} \end{aligned}$$

Experimental Results

Two expected observations....

Whole Probability Mass

is technically covered by a small set of reconciliations located close to α_{min}.

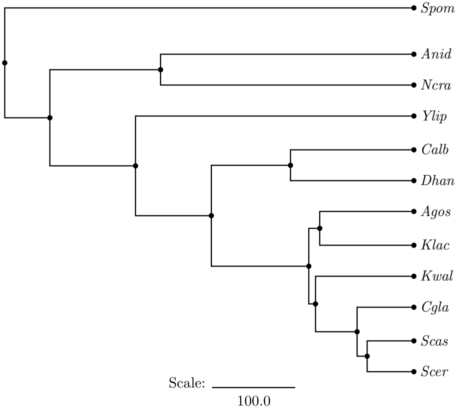
Approximation of the Most Probable Reconciliations

can easily and efficiently be computed.

Experimental Results

Input Data

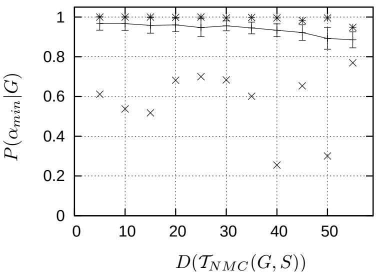
- 12 fungal genomes and 1278 gene family trees.
- Branch lengths computed by a Bayesian Framework.
- Branch rates estimated by an Expectation Maximization approach.

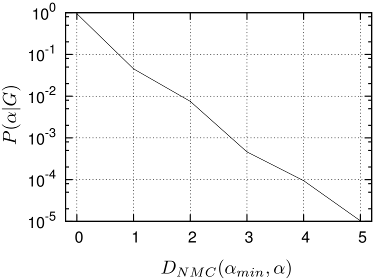


Experimental Results

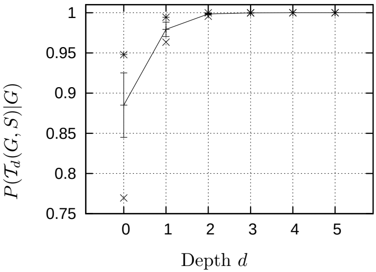
1278 Fungal gene trees

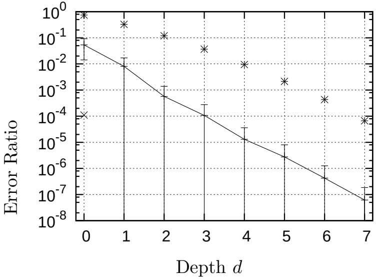
- In 1276 cases, α_{min} is the Most Probable Reconciliation.
- α_{min} covers most of the Probability Mass.
- The more a reconciliation is located close to α_{min}, the more it is probable.





Experimental Results





$$\text{Error Ratio} = 1 - \frac{\textit{approx}}{\textit{exact}}$$

Conclusion

Main Observations

1. Small # of reconciliations are needed to approximate probabilities.
2. The neighborhood of α_{min} “covers” the probability mass.
3. Similar results for synthetic gene trees generated with higher rates.

Future Work

1. Higher rates along S.
2. Reconciliation spaces where α_{min} is located far from α^*?
3. Reconciliation spaces with more than one peaks?
4. Similar Bayesian Framework for dup./loss rates probabilistic analysis.