

Inferring a Duplication, Speciation and Loss History from a Gene Tree

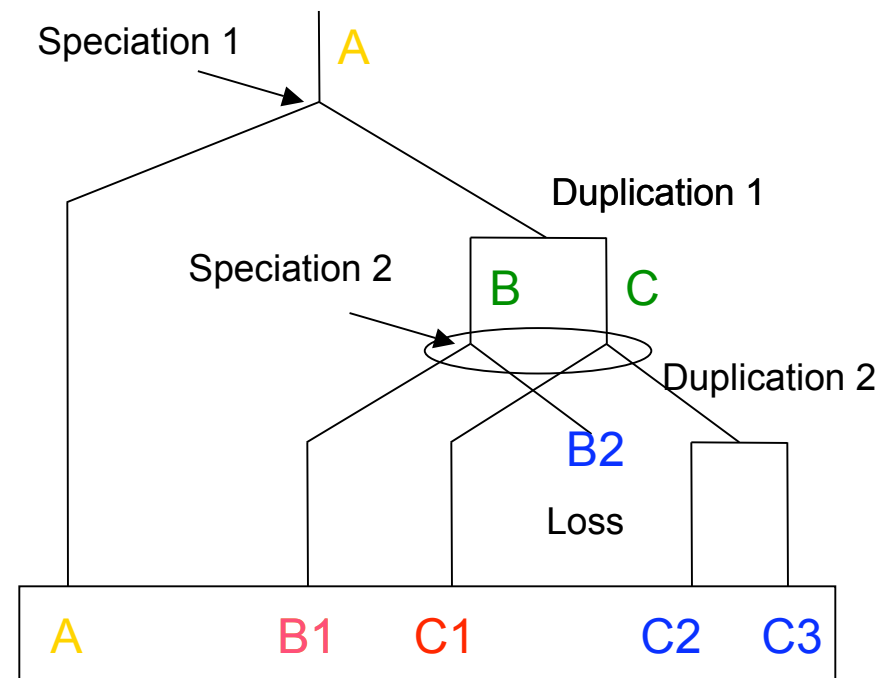
Cedric Chauve, Dept. Mathematics, Simon Fraser University
and Comparative Genomics Lab., UQAM

Nadia El-Mabrouk, DIRO, Université de Montréal

Jean-Philippe Doyon, DIRO, Université de Montréal

Evolution of gene families

- **Gene family:** gene present in current genomes having evolved from a single ancestral gene through duplication, speciation and loss events.
- **Duplications:** tandem, chromosomal fragment whole genome. Major evolutionary mechanism (Ohno, 1970): more than 65% of Arabidopsis genes belong to a gene family with duplicates.
- **Loss:** a copy of a gene is not used anymore or changes function, and as a result, its DNA sequence diverges rapidly from the sequences of other copies.



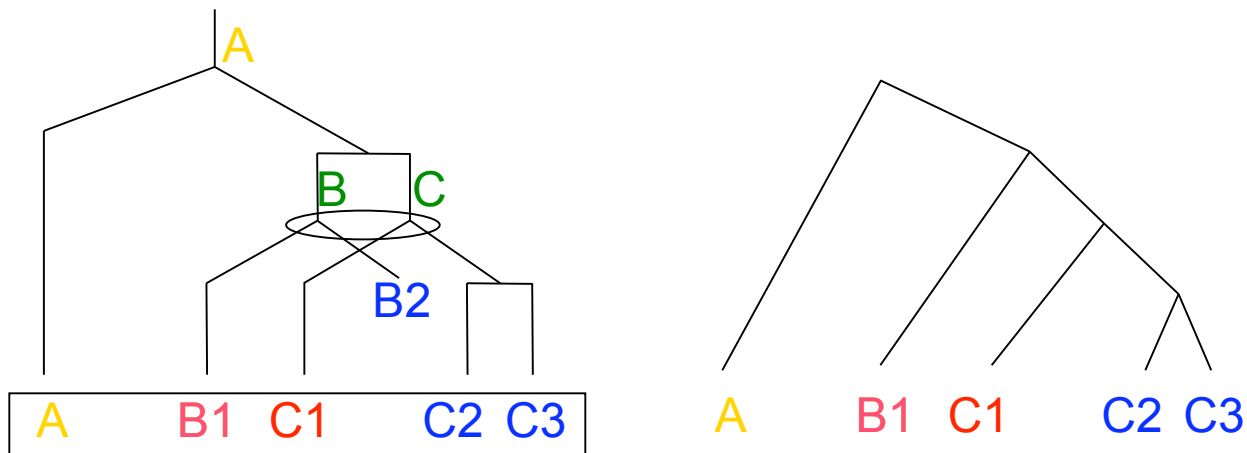
orthologs: (A;B1,C1,C2,C3) or (C1;C2,C3)

paralogs: (B1;C1,C2,C3) or (C2;C3)

Applications: understanding evolutionary mechanisms, functional annotation, phylogenomics and genome rearrangements studies

Analysis of gene families.

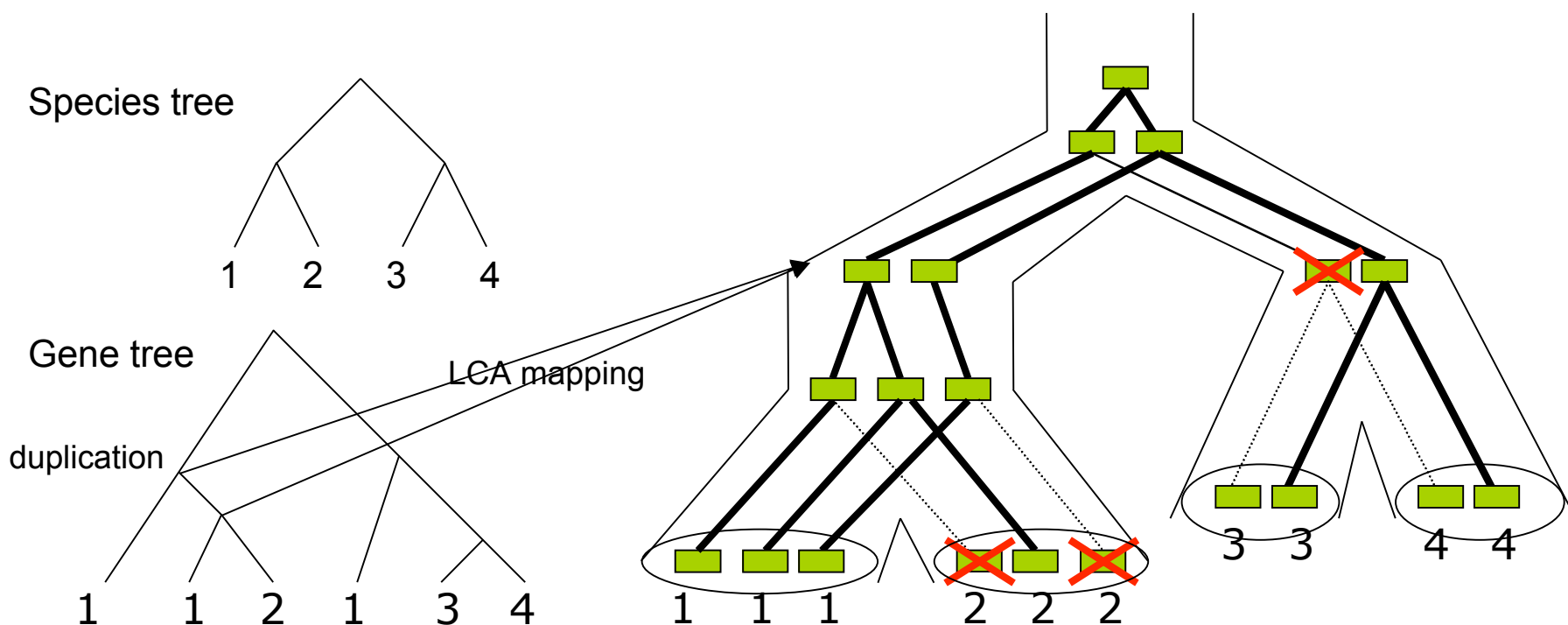
- First step: identification of gene families by sequence similarity
- Second step: given a gene family, computing a phylogenetic tree (called here the gene tree, and that we supposed fully resolved)



- General question: from the gene tree, can we recover the evolutionary history of this family (species tree, duplications, losses) ?

When the species tree is known

- Gene tree/species tree reconciliation (Page & Charleston 1997; Cotton & Page 2002): Embedding the gene tree into the species tree allows to identify possible duplication and loss events. Linear time algorithms (Zhang, 1995).



When the species tree is not known

- Given a (or a set of) gene tree(s), infer a species tree that minimizes:
 - the number of **duplications** (Duplication-cost model)
 - the number of **duplications + losses** (Mutation-cost model)

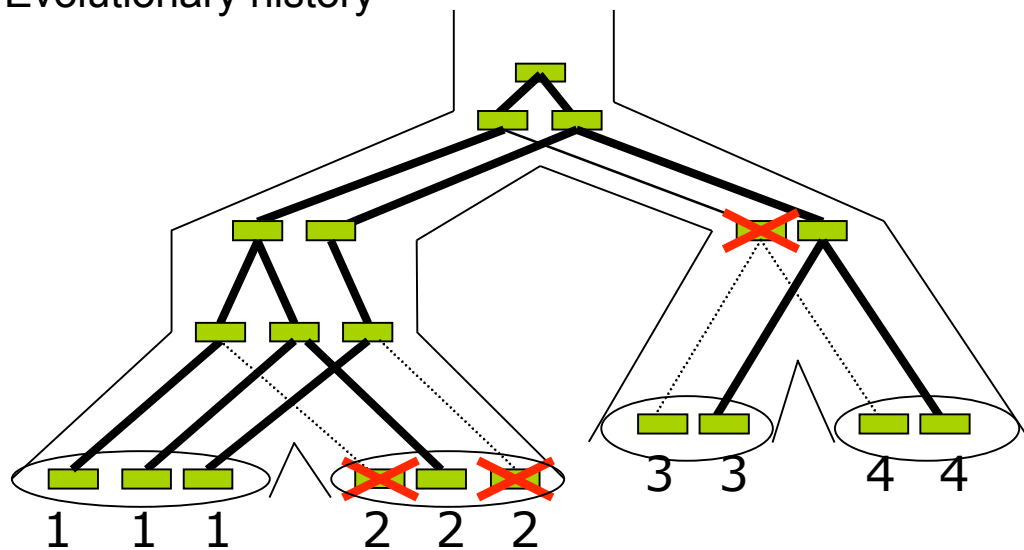
- Both problems are **NP-hard** but fixed-parameter tractable (Hallett and Lagergren 1996, Ma, Li and Zhang, 2000, Hallett and Lagergren 2000).

- **Here we are interested only in loss events:**
 - Given a gene tree G , can it be explained only by duplications and speciations (i.e. **with no loss**) ? Such trees are called **DS-trees** and have been introduced by Page, when the species tree is known, as reconciled trees.

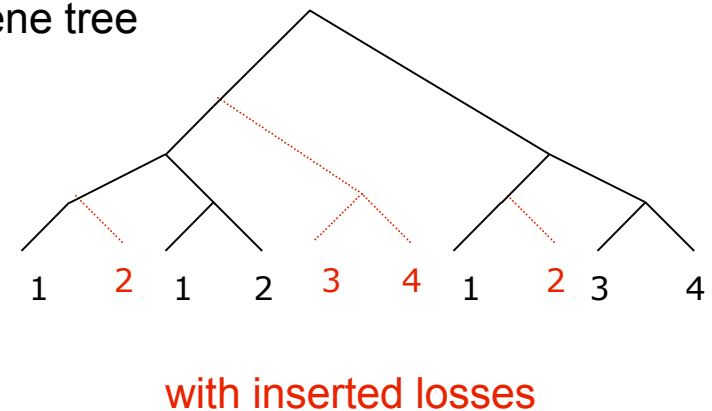
 - If G can not be explained without losses, what is the minimum number of losses that could have occurred ?

Rationale for looking at losses only

Evolutionary history

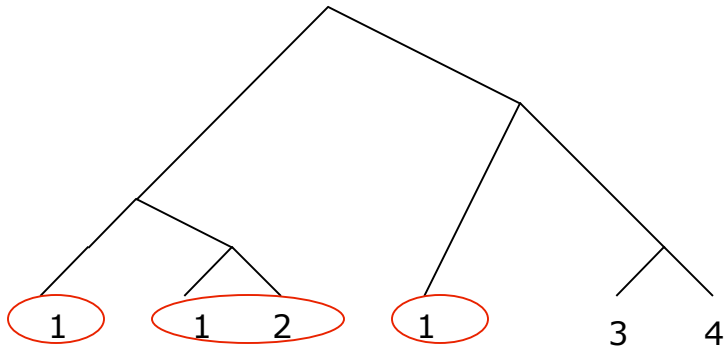


Gene tree

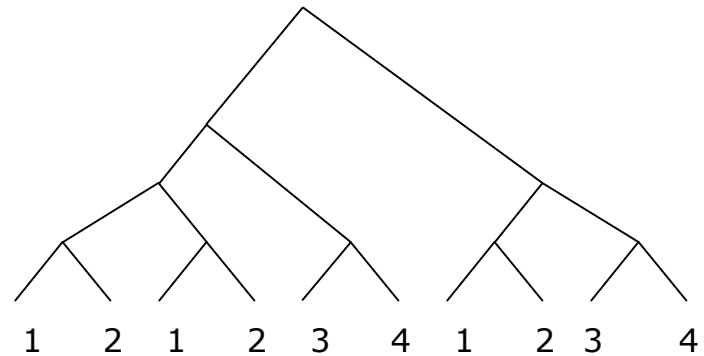


- Provided that the gene tree G is correct, adding the **true loss events** to G gives a **non-ambiguous evolutionary history**.
- Considering the **minimum** number of losses is a **natural combinatorial criterion**.

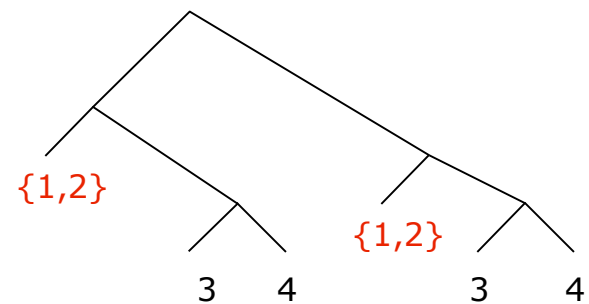
Is G a DS-tree? Bottom-up approach (1)



non DS-tree:
 $\{1,2\}$ is not a **valid cherry**:
occurrences of 1 are not coupled
to a 2.



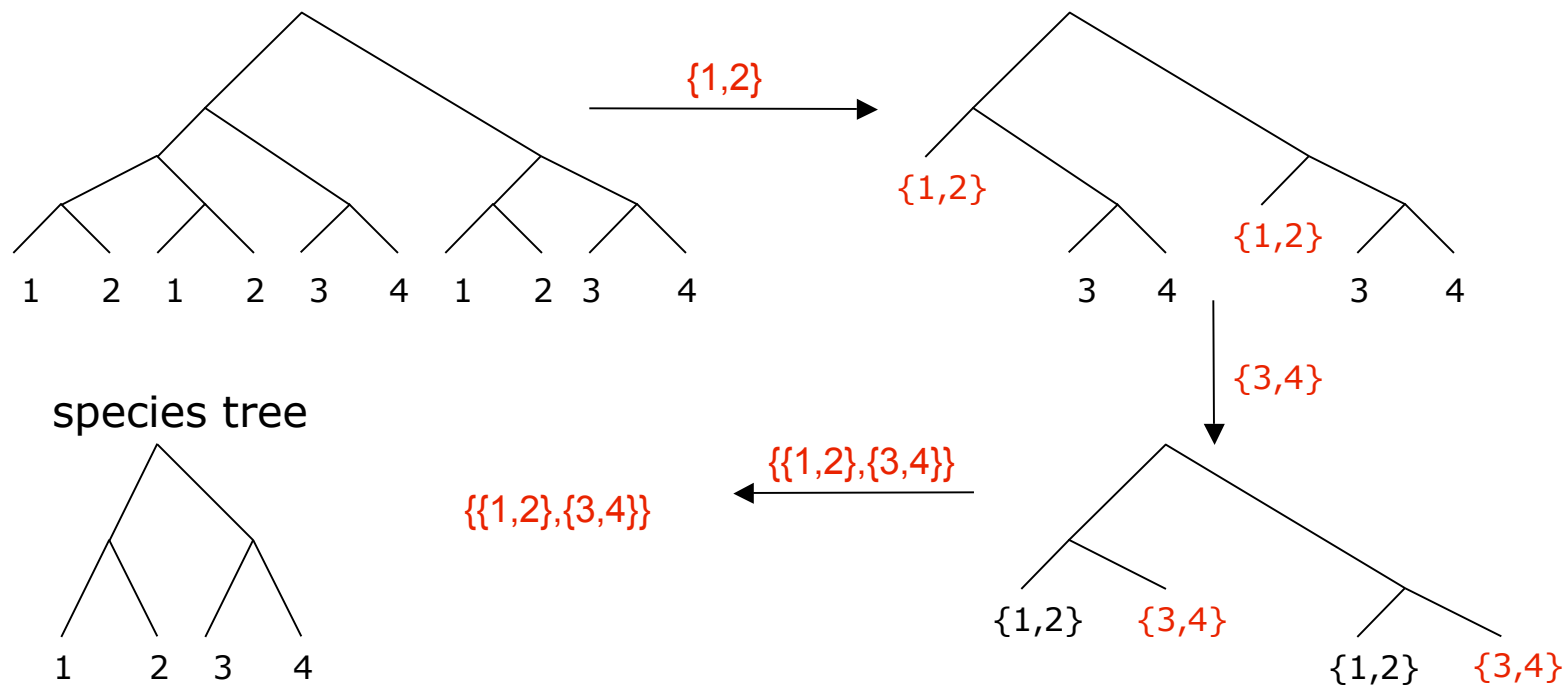
DS-tree:
 $\{1,2\}$ is a **valid cherry**, and can
be **contracted**.



Is G a DS-tree? Bottom-up approach (2)

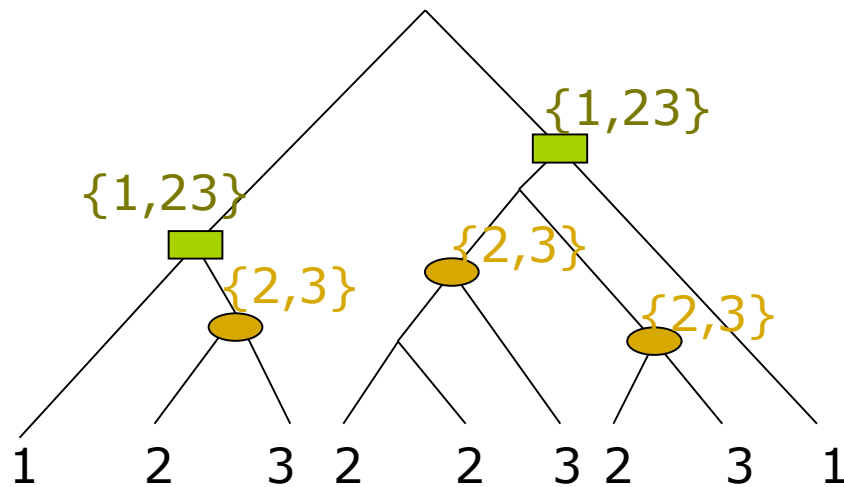
Theorem.

1. G is a DS-tree iff it has a single leaf or it has a **valid cherry** $\{i,j\}$ and the tree obtained by **contracting** $\{i,j\}$ is a **DS-tree**.
2. If G is a DS-tree, it is compatible with a single species tree.



This process can be implemented to run in **linear time and space**.

Is G a DS-tree? Top-down approach



Theorem.
G is a DS-tree iff it is DS-valid.

- **DS-Valid nodes:** nodes such that the intersection between the leaf sets of their two children is empty.
- **Border:** F forest, V, the set of higher valid nodes in F, is a border for F iff the nodes of V cover all leaves of F and all leaf sets partitions induced by the children of the nodes of V are identical.
- **F is DS-valid** iff it is a set of trees with a unique leaf label or it has a border and the two forests induced by the children of the nodes of this border are DS-valid.

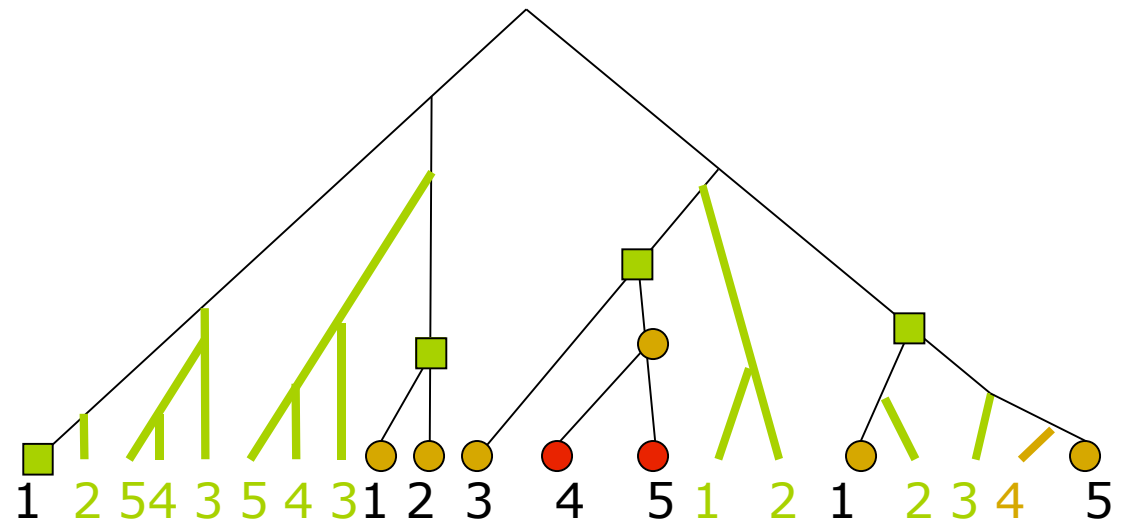
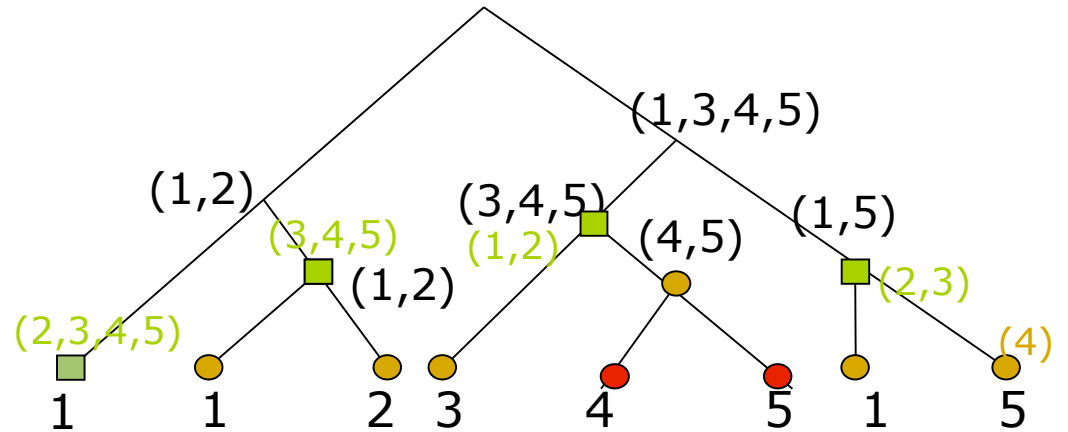
Minimizing losses

- Let G be a gene tree that is not a DS-tree.
- Two equivalent problems:
 - What is the minimum number of losses (subtrees) to insert in G to transform it into a DS-tree ?
 - What is a species tree S such that the reconciliation (LCA mapping) between G and S induces a minimum number of losses ?
- Our results:
 - A **heuristic** that transforms G into a DS-tree, in time and space $O(g \cdot n)$, where g is the number of genomes and n the number of genes.
 - A **branch-and-bound** algorithm that finds the exact solution.
 - An **experimental study** on a dataset of **plant genes**, with 7 genomes and 577 gene trees.

Heuristic 2: inserting losses ...

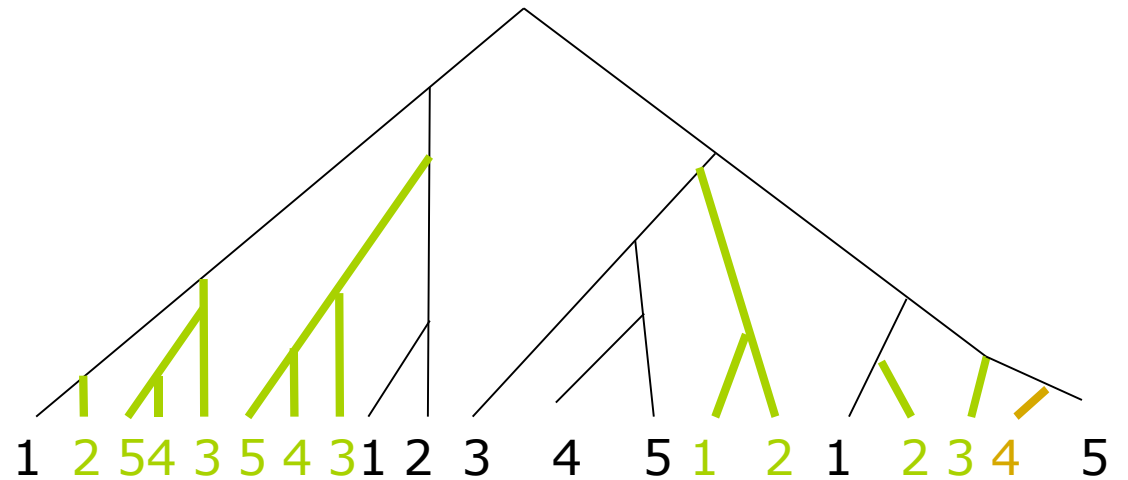
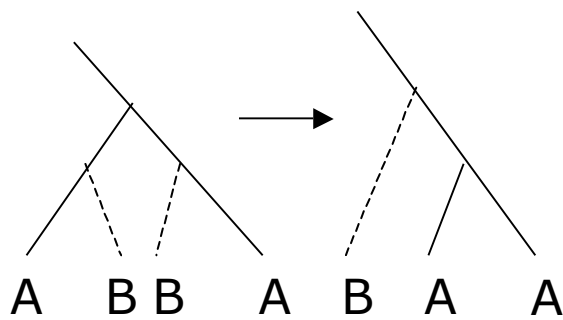
... to get a DS-tree

1. Start from the lowest level of nodes that need losses insertions: ●
2. For each label L of this level, let P(L) be a phylogeny of this leaf set.
3. For every node of the current level, labelled L, complete its subtree according to P(L)
4. Iterate for the next level up: ■

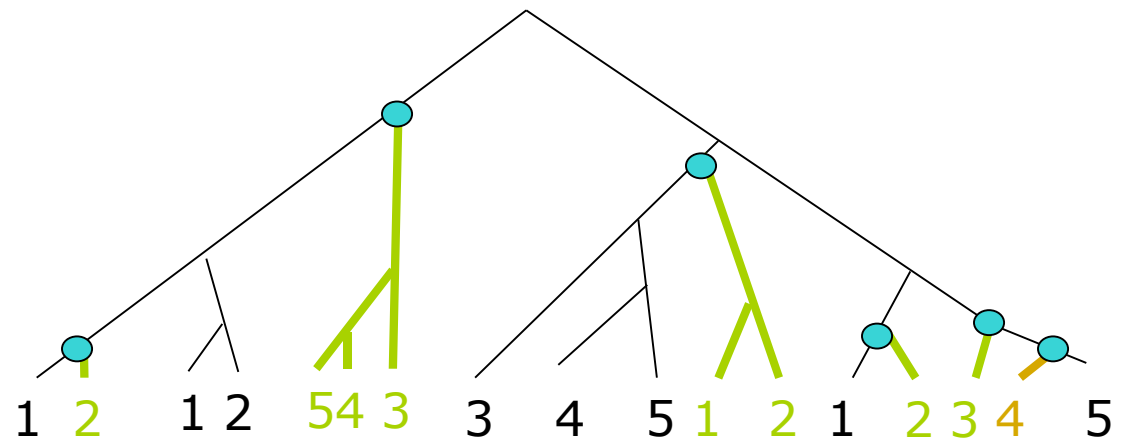
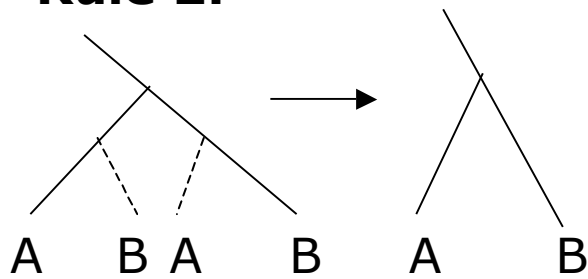


Heuristic 3: factorizing losses

Rule 1:



Rule 2:



Branch-and-Bound algorithm

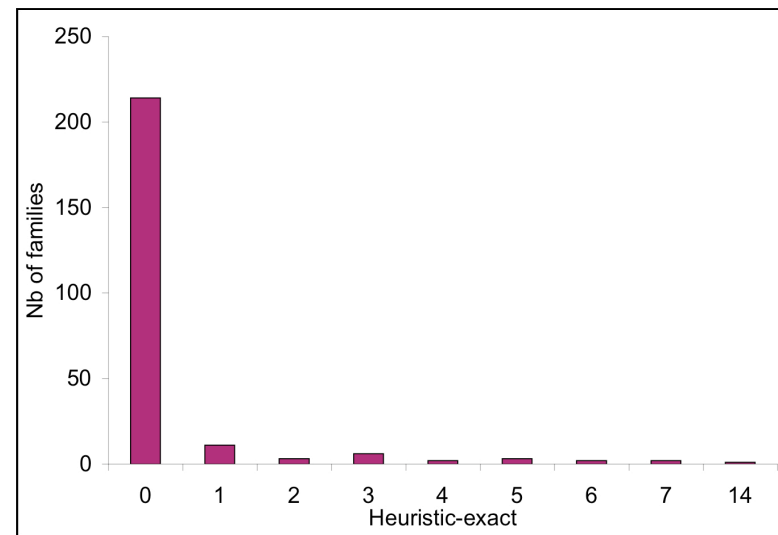
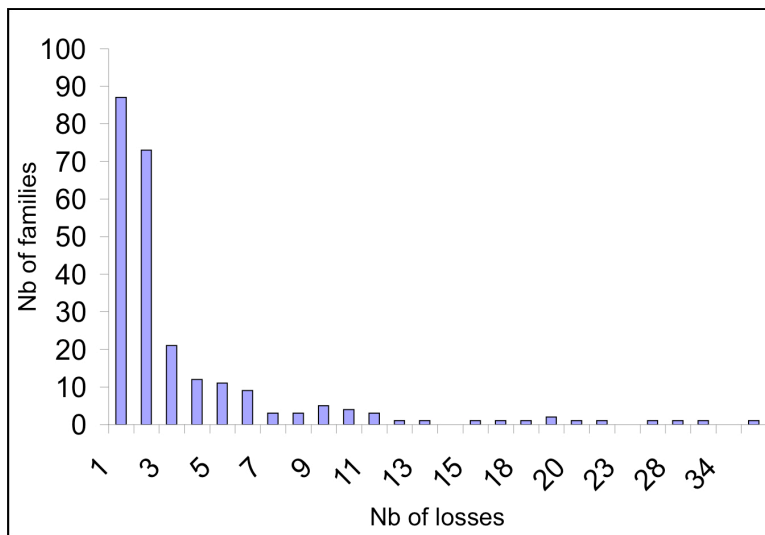
- ❑ Based on the **bottom-up** approach to decide if a gene tree is a DS-tree.
- ❑ Motivation: the number of possible **bipartitions** at the root of species tree is **exponential** in the number of species, while the number of possible **cherries** is **quadratic**.
- ❑ General principle
 - the current incomplete species tree is a forest (all trees are leaves at the beginning);
 - during each step of the branch-and-bound process, two trees of the current forest are joined in a single tree, and the number of losses is updated;
 - updating the number of losses is done without computing the reconciliation but in time proportional to the number of losses;
 - simple and efficient (optimal ?) algorithm, but we did not find in the literature any algorithmical study of branch-and-bound algorithms for the duplication, duplication+loss or loss models.

Experimental results (1)

- 577 gene trees from a study of angiosperm genomes, with an outgroup (Sanderson & McMahon, 2007):
 - all gene trees are phylogenetically informative
 - 59 contain genes from all 7 genomes
 - gene families were computed by single-linkage clustering
 - gene trees were computed using PAUP and Maximum Likelihood
- Results on DS-trees:
 - 333 gene trees are DS-trees
 - many of them, but not all, with few genes and/or duplication events:
 - 89 DS-trees (32 nonDS) with 4 genes and 3 species
 - 8 DS-trees (14 nonDS) with 6 genes and 3 species
 - 17 DS-trees (10 non DS) with 6 genes and 4 species
 - 14 DS-trees (0 nonDS) with 6 genes and 5 species
 - 1 DS-tree (2 nonDS) with 13 genes and 5 species
 - 7 DS-trees (over 59) with 7 species, the largest with 11 genes

Experimental results (2)

- 244 gene trees on 577 are not DS-trees:
 - we used both the heuristic and the branch-and-bound to assess how many gene losses are needed to propose an evolutionary history;
 - many of them can be explained with few losses;
 - the heuristic performs very well in general;
 - the branch-and-bound algorithm proposes in most of the cases a single optimal species tree (179 times for 244 trees).



Experimental results (3)

- The 159 gene trees that contained a gene from the outgroup were gathered under a single non-binary root and similar experiments (heuristic and branch-and-bound) were performed.
- Quality of the heuristic (number of loss events):
 - Heuristic: 2200
 - Branch-and-bound: 1502
- Inferred species tree:
 - Branch-and-bound: very close from the accepted tree (1 branch swap)
 - Heuristic: less good, with one minor branch swap and one major branch swap.

Conclusion

- Results:
 - The concept of Duplication-speciation history occurs naturally in the study of gene families.
 - It leads to the natural problem of minimizing the number of losses to explain the evolution of a gene family.
 - We propose algorithms for the problems of deciding if a gene tree is a DS-tree and minimizing the number of losses.
 - Our algorithms perform quite well on a small dataset.
- Perspectives and work in progress:
 - complexity of the problem of minimizing the number of losses (probably NP-hard);
 - algorithmical improvements (choice of the phylogeny to complete nodes in the heuristic);
 - comparative study, both theoretical and experimental, of the three available combinatorial criteria (dup., dup+losses, losses);
 - extension to non-binary gene trees.