Inferring a duplication and speciation history from a gene tree [2]

Cedric Chauve[1], **Jean-Philippe Doyon**[2], **Nadia El-Mabrouk**[2]

(1) LaCIM and CGL, Université du Québec à Montréal, Canada (2) DIRO, Université de Montréal

INTRODUCTION

Gene families are composed of homologous genes that share a common ancestor. Each is the result of an (unknown) **evolution scenario** involving gene duplication, speciation and gene loss.

Importance of knowing the evolutionary scenario of each gene family

1. to identify **orthologous** and **paralogous** genes: pair of genes separated respectively by a speciation and a duplication event;
2. to annotate genes: orthologs have, in general, similar functions;
3. to map genes between genomes (needed for gene order and rearrangement analyses).

To infer gene family evolutionary scenarios, we construct the gene tree T and compare it with the species tree S to deduce gene duplication and loss events. This comparison is called "gene tree / species tree reconciliation" [3], can be done by parsimony [5] or probabilistic [1] methods and is based on the "Last Common Ancestor" mapping of the nodes of T to the nodes of S.

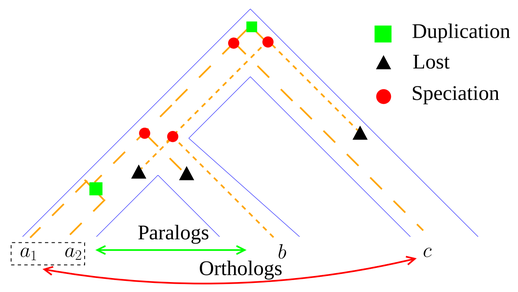


Fig. 1. A gene family evolutionary scenario is represented by the reconciliation of the gene tree (dotted lines) and the species tree (blue lines).

"WHAT DO WE DO IF WE DON'T KNOW THE SPECIES TREE?"

Given a gene tree T, we study the following two problems:

1. **DS-recognition problem**: can T be explained by a history involving only duplication and speciation events? Such a gene tree is called a "DS-tree".
2. **Gene losses minimisation problem**: what is the minimum number of losses needed to be inserted in T to transform it into a DS-tree?

Contributions

1. Bottom-Up and Top-Down approaches defining DS-trees.
2. A **linear time and space** algorithm (Bottom-Up approach) for the *DS-recognition problem*. It also computes the **unique species tree** that is consistent with the DS-history.
3. A heuristic for the *Gene losses minimisation problem*.

TWO APPROACHES TO CHARACTERIZE DS-TREES

Let $\mathcal{G} = \{1, \ldots, g\}$ be a set of g species labels, and T be a gene tree whose leaf labels belong to $\mathcal{G}$. For an internal vertex x of T, we note by $L(x) \subseteq \mathcal{G}$ the label set induced by the leaves of the tree T_x rooted at x, and by x' its sibling.

Bottom-Up approach (leads to the DS-recognition algorithm)

T is a DS-tree on $\mathcal{G}$ iff. it respects the following conditions.

1. Let x be an internal node of T s.t. $L(x) = i \in \mathcal{G}$ and $L(x') = j \in \mathcal{G} \setminus \{i\}$. Check that for all internal nodes y of T, $L(y) = i$ iff. $L(y') = j$, and replace their father by leaves labelled by the new species $g + 1$.
2. Check that this new tree is DS on $\mathcal{G} \setminus \{i, j\} \cup \{g + 1\}$. Repeat the process until $|\mathcal{G}| = 1$.

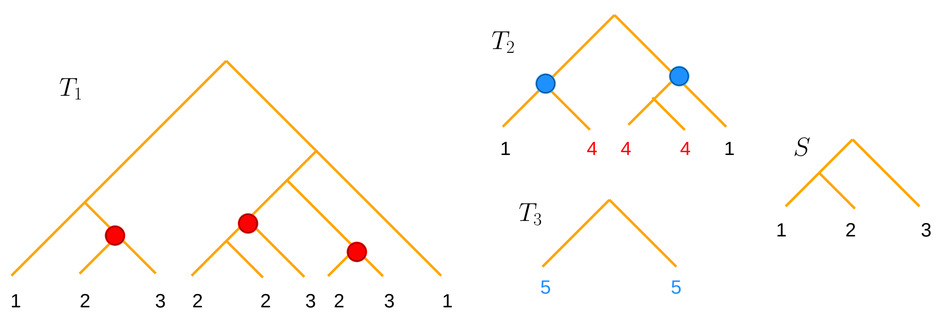


Fig. 2. T_1, T_2 and T_3 are gene trees (leaf genes are labelled by the corresponding species). In T_1 (resp. T_2), each red (resp. blue) node is replaced by a leaf labelled by the (new) species 4 (resp. 5) in T_2 (resp. T_3). S is the corresponding species tree.

Top-Down approach

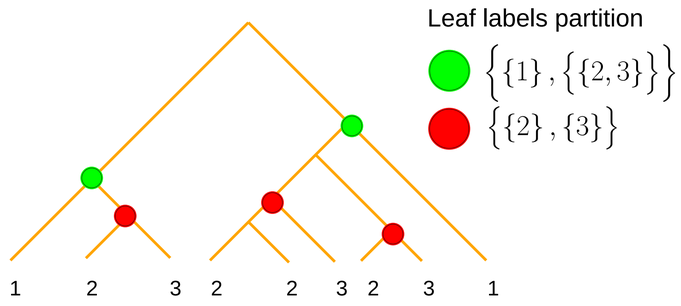


Fig. 3. In the 1th (2th) iteration, the internal green (resp. red) nodes uniquely cover each of the considered leaves and they all have the same partition of the corresponding leaf labels. The iteration process continue until there is no internal node left.

HEURISTIC FOR THE *Gene losses minimisation problem*

The heuristic allowing to obtain an upper bound on the *Gene losses minimisation problem* is based on the Top-Down approach and is decomposed in the following three steps.

1. Recursively modify the label set of the vertices of T (beginning at the root) s.t. any pair of vertices from the same recursive level have disjoint or equal label set.
2. Consider successively each level beginning with the last one and perform the following steps: 1) partition the vertices of the current level according to their new label set; 2) for each partition, arbitrary choose a phylogeny for the common label (species) set and perform the losses insertions leading to this set.
3. Finally, it is possible to reduce the number of insertions by applying the factorization rules.

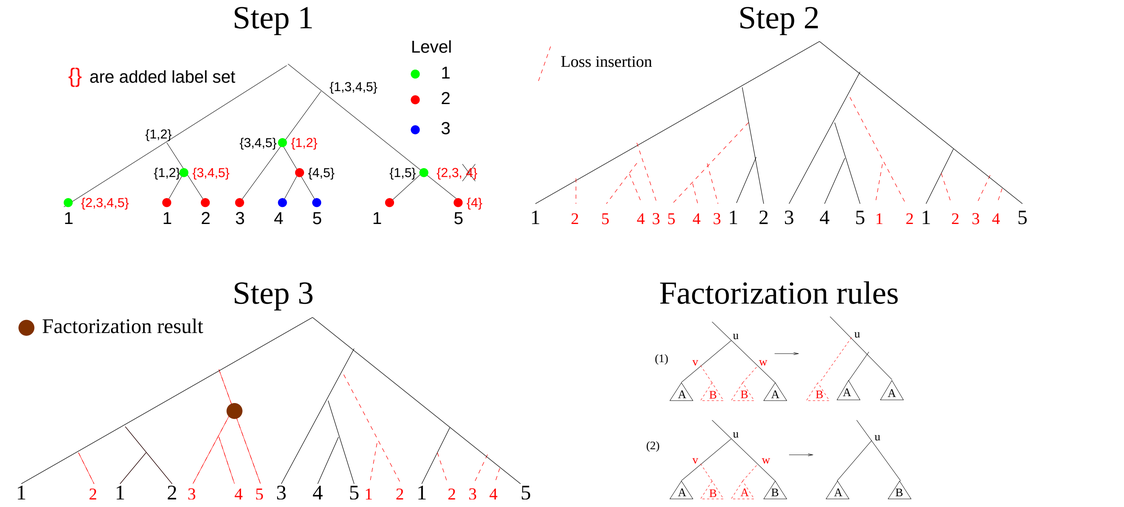


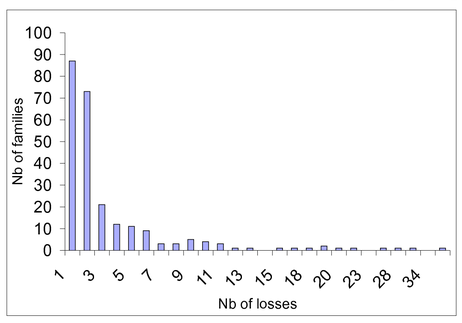
Fig. 4. An illustration of the heuristic. For the factorization rules, u is a vertex presents in the original tree.

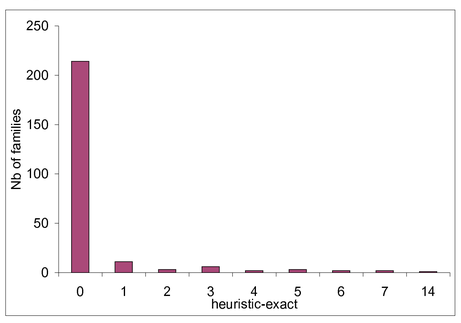
The **time complexity** of the whole algorithm is in $O(gn)$, where g and n are respectively the number of genomes and the size of the gene tree.

EXPERIMENTAL RESULTS

First, among 577 gene families from a study [4] of the phylogeny of seven angiosperm genomes from EST data, 333 were found to be DS-tree (most of these exhibit few gene duplications). Second, we applied to the remaining 244 gene trees our heuristic to compute an upper bound on the minimum number of gene losses needed to explain the observed gene tree. The results are summarized in the three figures below, where each of them give the distribution of the 244 gene families according to three different parameters.

- **Left**: the number of gene losses inferred by our heuristic. Many gene families can be explained with few gene losses **!!**
- **Right**: the difference between the number of losses of the heuristic solution and of the optimal one (computed by a branch-and-bound algorithm). Our heuristic performs well **!!**
- **Bottom**: the number of species trees inducing the minimum (optimal) number of gene losses. In most cases, there is a unique optimal species tree **!!**





Number of species trees	1	2	3	4	5	6	7	13	15
Number of families	179	16	34	2	6	3	1	2	1

CONCLUSION

One of the main objectives of the whole project is to propose one or more credible species phylogenies so that they can be used as a startup for phylogenetic inference methods. When we considered our data set, we used the supertree methods on the whole set of species phylogenies computed by the DS-recognition algorithm to build a species supertree. However, we observed that this phylogeny has some unresolved branches.

The next step of the project is to incorporate all gene family trees into one big tree (its polytomy root will represents one big duplication), and to apply our algorithms on the latter. We hope that this will solve the lack of resolution.

References

[1] L. Arvestad, A.C. Berglund, J. Lagergren, and B. Sennblad. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In *RECOMB 2004*.

[2] C. Chauve, J.P. Doyon, and N. El-Mabrouk. Inferring a duplication, speciation and loss history from a gene tree. *The Fifth RECOMB Comparative Genomics Satellite Workshop*, 2007 (To appear).

[3] J. A. Cotton and R. D. Page. Going nuclear: gene family evolution and vertebrate phylogeny reconciled. *Proc Biol Sci*, 269(1500):1555–61, Aug 2002.

[4] M. J. Sanderson and M. M. McMahon. Inferring angiosperm phylogeny from est data with widespread gene duplication. *BMC Evol Biol*, 7 Suppl 1:S3, 2007.

[5] C. M. Zmasek and S. R. Eddy. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, 17(9):821–828, Sep 2001.

Address for correspondence: Département d'Informatique, UQÀM, CP 8888, succ. centre-ville, Montréal (QC), H3C 3P8