

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 infer **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).

Protocole to infer gene family evolutionary scenarios

1. Infer the gene family phylogenetic tree G .
2. Compare G to the species tree S to deduce gene duplication and loss events.

This comparison is called “gene tree / species tree reconciliation” [2], can be done by parsimony [4] or probabilistic [1] methods and is based on the “Last Common Ancestor” (LCA) mapping of the nodes of G to the nodes of S .

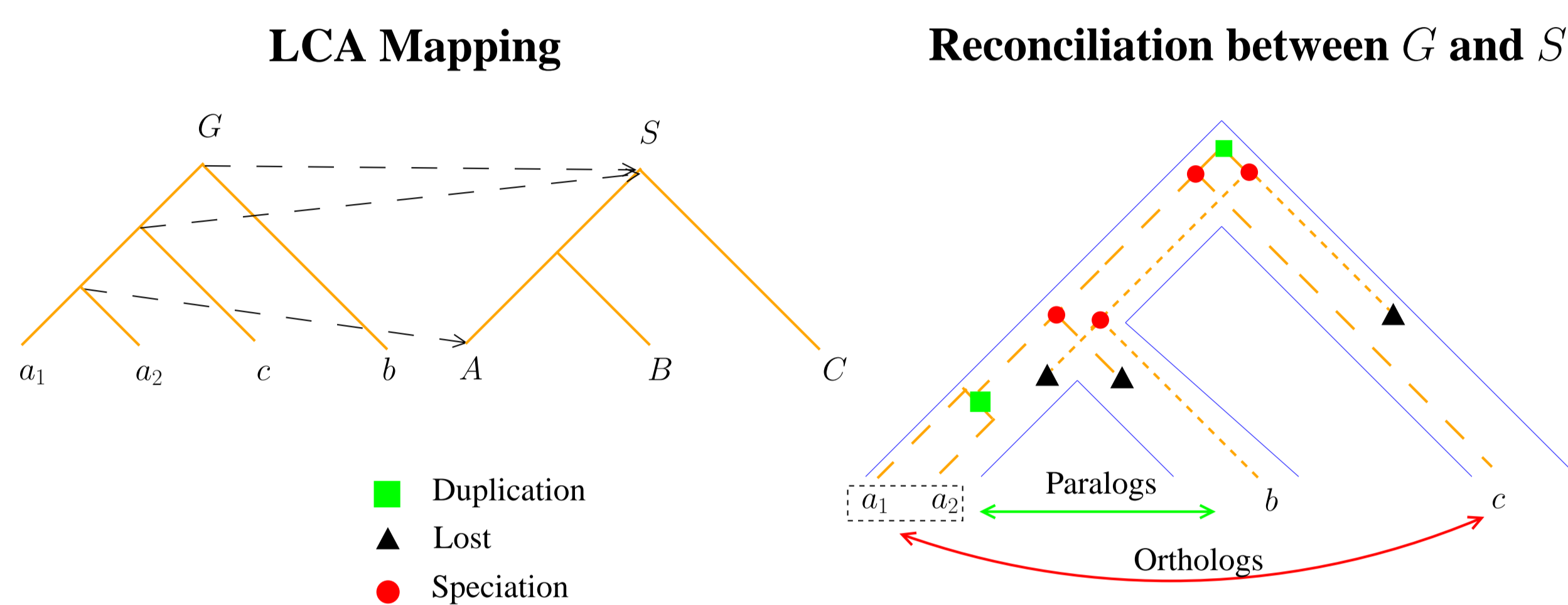


Figure 1: LCA allows to infer reconciliation between gene tree G and species tree S .

“WHAT DO WE DO IF WE DON’T KNOW THE SPECIES TREE?”

Problem

1. Given a gene tree G , can it be explained by a history involving only speciation and duplication events (without gene loss)?
2. Such a gene tree G is called a “Duplication/Speciation tree” (“DS-tree”).

Contributions

1. If G is a DS-tree, the species tree S consistent with G is unique.
2. An algorithm that decides if G is a DS-tree and, if so, compute the corresponding species tree. It can be implemented to run in $O(n)$ time and space, where n is the number of nodes of G .

RECOGNIZING DS-TREES

Let $[g] = \{1, \dots, g\}$ be a set of g species labels, and G be a gene tree whose leaf labels belong to $[g]$. We note $L(G_u) \subseteq [g]$ the label set induced by the leaves of the rooted tree G_u , for $u \in V(G)$. For an internal node $u \in V(G)$, its children are denoted u_1 and u_2 .

Recursive definition of a DS-tree

A subset $\{i, j\} \subseteq [g]$ is called a **cherry** $\iff \exists u \in V(G)$ such that $L(G_u) = \{i, j\}$.

1. $\{i, j\}$ is a “DS-valid” cherry $\iff \forall u \in V(G)$ such that $L(G_{u_1}) = \{i\}$ (resp. $\{j\}$) and $L(G_{u_2}) \neq \{i\}$ (resp. $\{j\}$) implies that $L(G_{u_2}) = \{j\}$ (resp. $\{i\}$).
2. If $\{i, j\}$ is a DS-valid cherry, we denote by $c(G, i, j)$ the gene tree whose nodes $u \in V(G)$, such that $L(G_u) = \{i, j\}$, are contracted and labelled by a new label representing $\{i, j\}$.

Theorem. G is a DS-tree on $[g]$ \iff either $g = 1$, or any cherry $\{i, j\}$ is DS-valid for G and $c(G, i, j)$ is a DS-tree on $[g] \setminus \{i, j\} \cup \{g+1\}$.

DS-tree recognition algorithm

Iteratively, consider a cherry $\{i, j\} \subseteq [g]$ and proceed with these 3 main steps:

1. If $\{i, j\}$ is not DS-valid, return FALSE; continue otherwise.
2. Contract all the occurrences of the cherry $\{i, j\}$: $G \leftarrow c(G, i, j)$.
3. Update the species tree with the current cherry.

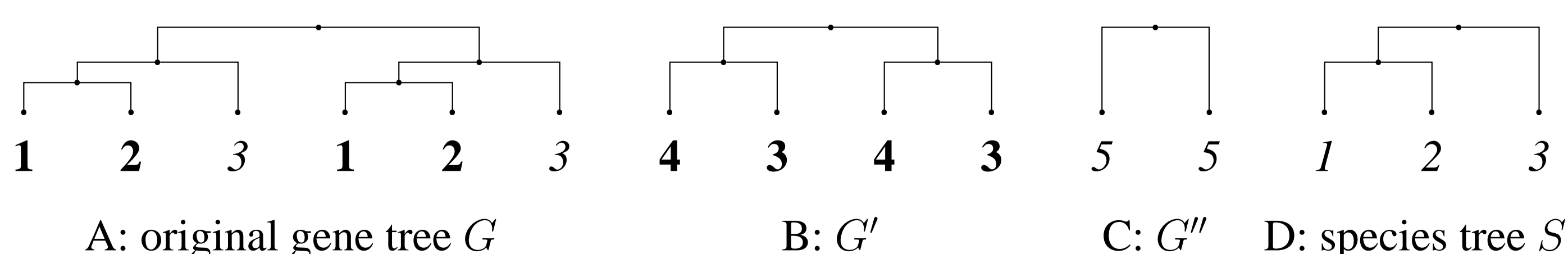


Figure 2: A and B: $\{1, 2\}$ and $\{4, 3\}$ are respectively the DS-valid cherries contracted; C: $\{5, 5\}$ is a “non-canonical” cherry; D: species tree consistent with G .

EXPERIMENTAL RESULTS

We consider 4 yeast species (*Candida glabrata*, *Yarrowia lipolytica*, *Kluyveromyces lactis* and *Debaryomyces hansenii*) that could be related according to 3 alternative topologies, and whose species tree is known.

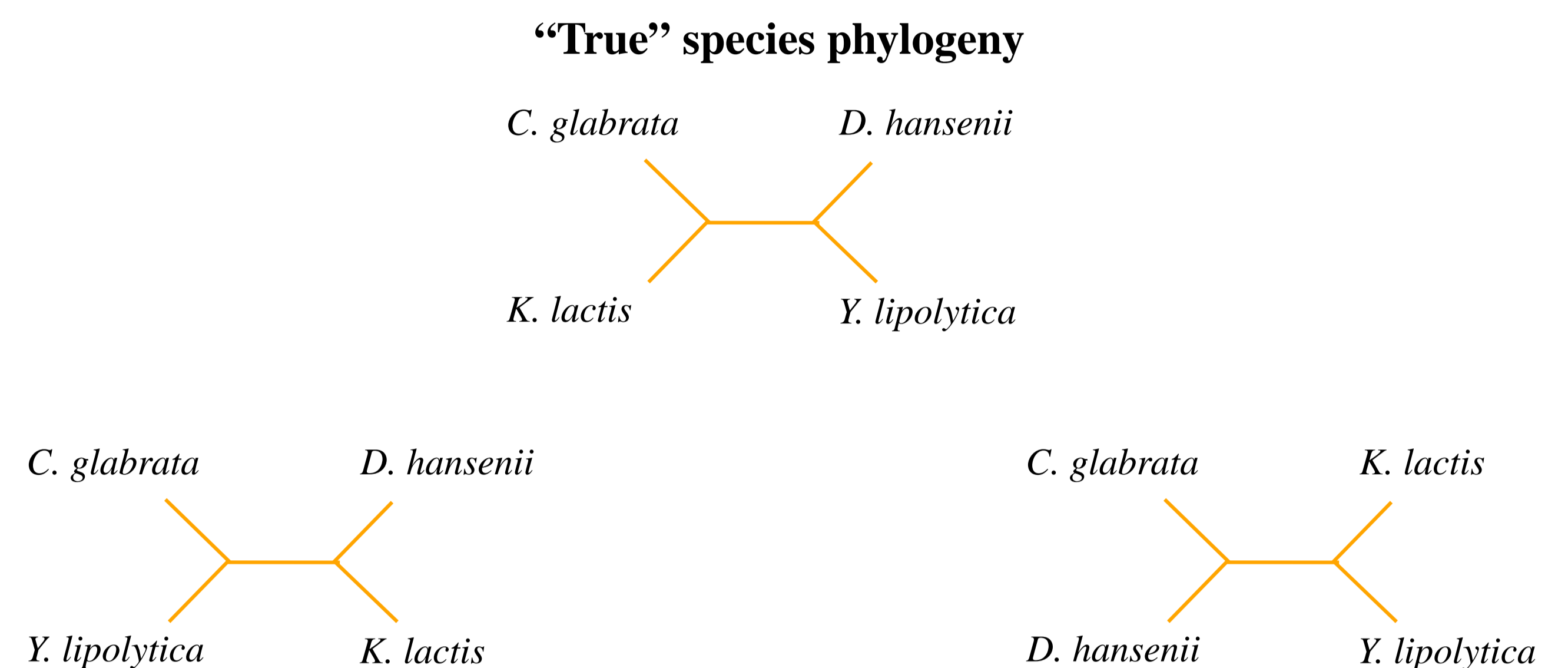


Figure 3: Three alternative possible topologies.

Gene family clustering analysis

Starting from the protein sequences of all genes, we applied OrthoMCL [3] to cluster them into gene families. All 24 165 genes were clustered into 5140 families from which 3334 were universal and, among those, 2902 were unique. A gene family is said to be

1. **universal** when all species are present;
2. **unique** when there is one and only one gene per species.

DS-tree recognition analysis

1. We constructed a phylogeny tree for each family with Phylml (JTT+F+ Γ +I model), and calculated statistical supports for each branch based on 100 bootstrap replicates.
2. We applied our algorithm to each family and classified the results according to 7 different bootstrap threshold values.

Bootstrap threshold	Nb. of trees		Nb. of DS-trees non-unique	... whose species tree is S	
	Unique	non-unique		Unique	non-unique
100	1404	63	60	1404	60
95	561	61	52	557	51
90	229	32	24	224	24
85	132	26	13	125	12
80	104	22	14	97	12
75	82	22	11	66	11
70	78	13	6	65	6
Total	2590	239	180	2538	176

Table 1. Results for all universal families. The number of families where the gene tree is a DS-tree is only shown for the non-unique families, given that a unique tree family is always a DS-tree. The two last columns give the number of those DS-trees whose species tree inferred is S , the “true” phylogeny.

CONCLUSION

We have developed an algorithm to verify if the evolutionary scenario of a given gene family can be explained only by duplication and speciation events (without gene losses).

In our yeast dataset, (1) most gene families can be explained without gene loss and (2) the species tree inferred from most of them is consistent with the “true” species tree. Thus, our algorithm can be useful to infer a valuable species phylogeny.

Gene families that **cannot be explained** without gene loss are frequent in eukaryotic evolution. The next question would thus be to infer the most parsimonious number of gene loss insertions that is needed to transform a non DS-tree into a DS-tree.

Because **lack of resolution** is very common in phylogenetic analysis, an important improvement to the algorithm would be to take the unresolved branches into account. This can be done by allowing polytomies, e.g., nodes with more than two children, in the input gene trees.

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] 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.
- [3] L. Li, C.J. Stoeckert, and D.S. Roos. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.*, 13(9):2178–2189, 2003.
- [4] 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.