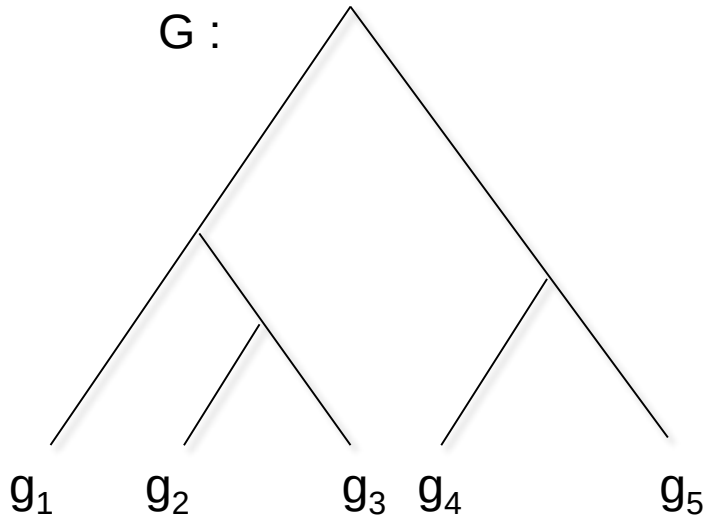
Error Detection and Correction of Gene Trees Using Gene Order

Manuel Lafond, Krister M. Swenson and Nadia El-Mabrouk

Université de Montréal

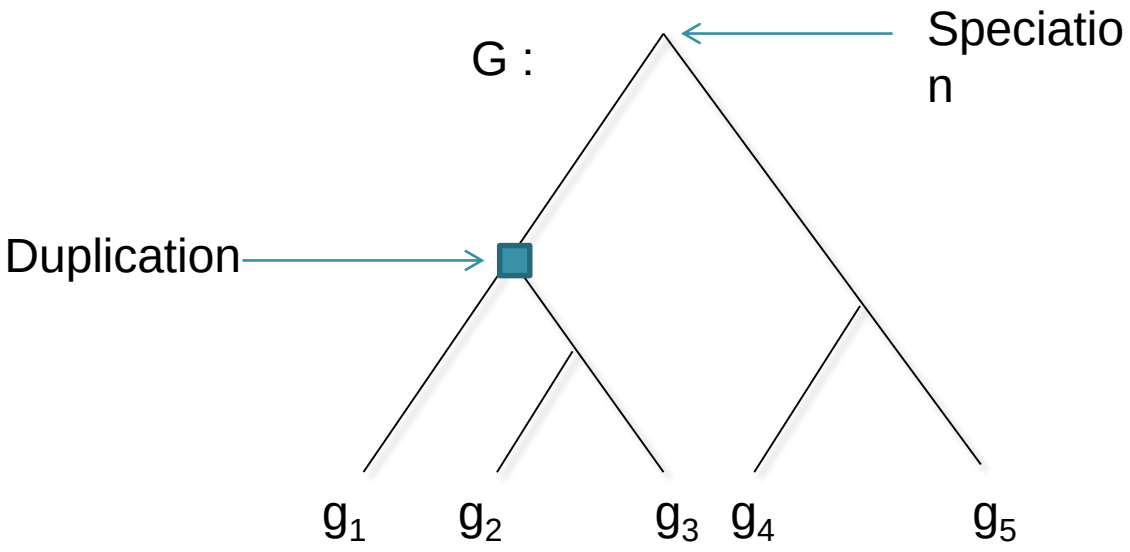
Introduction

- **Gene trees** reflect the evolutionary history of a family of **homologous** genes
  - Genes that all descend from a common ancestor



Introduction

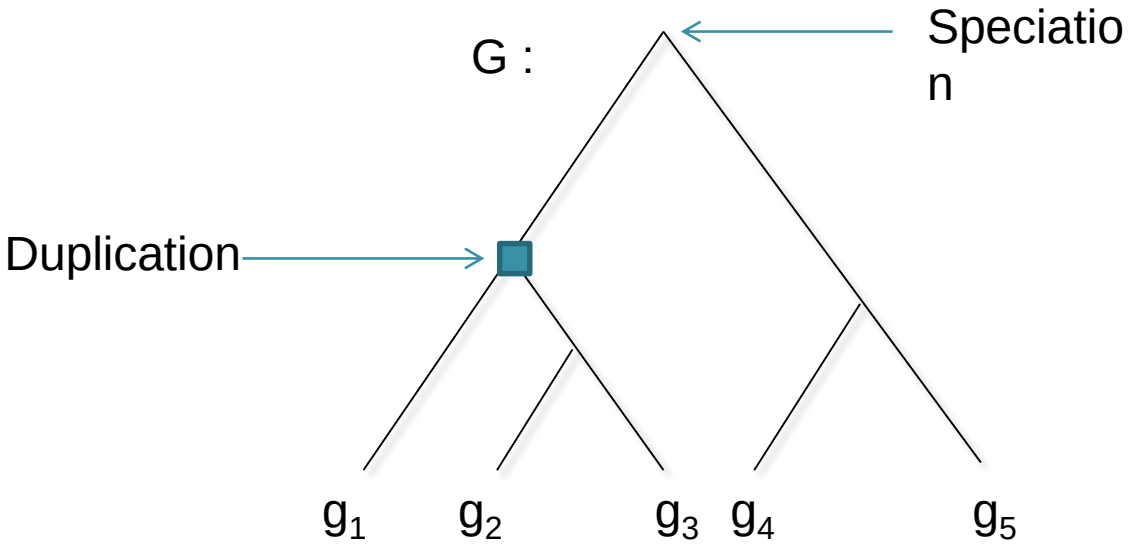
- Ancestral genes may have undergone **speciation** or **duplication**



Introduction

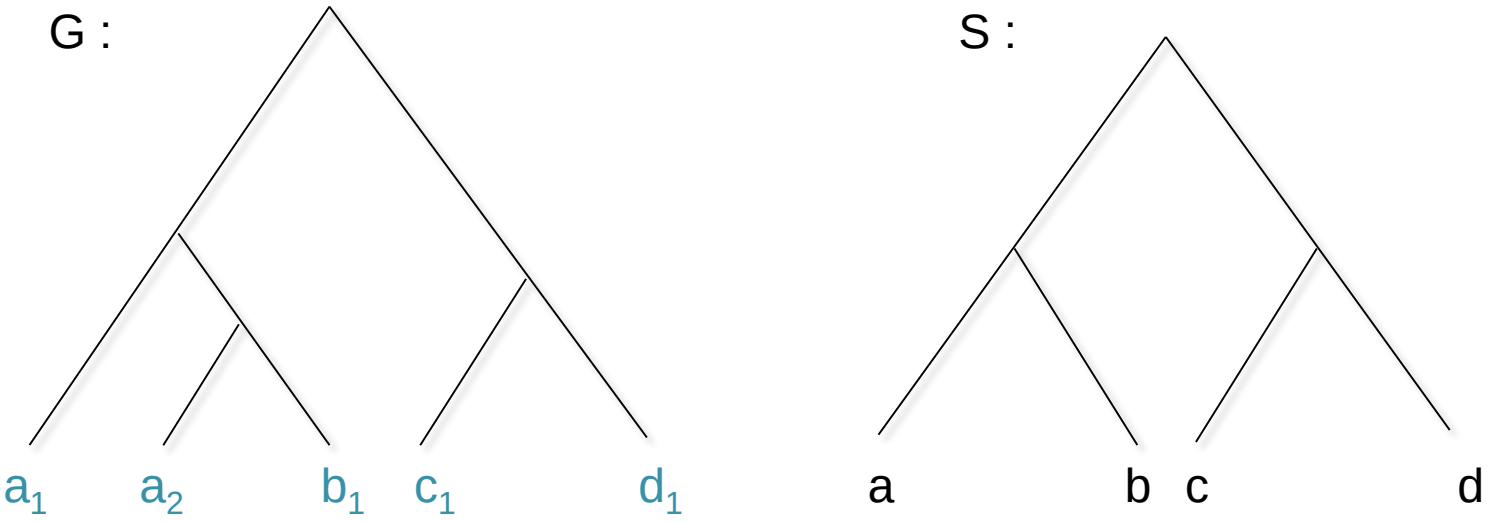
- Modern genes relationships
  - **Orthologs** : LCA is a speciation
    - g_1, g_5 are orthologs
  - **Paralogs** : LCA is a duplication
    - g_1, g_3 are paralogs

(LCA = Lowest Common Ancestor)



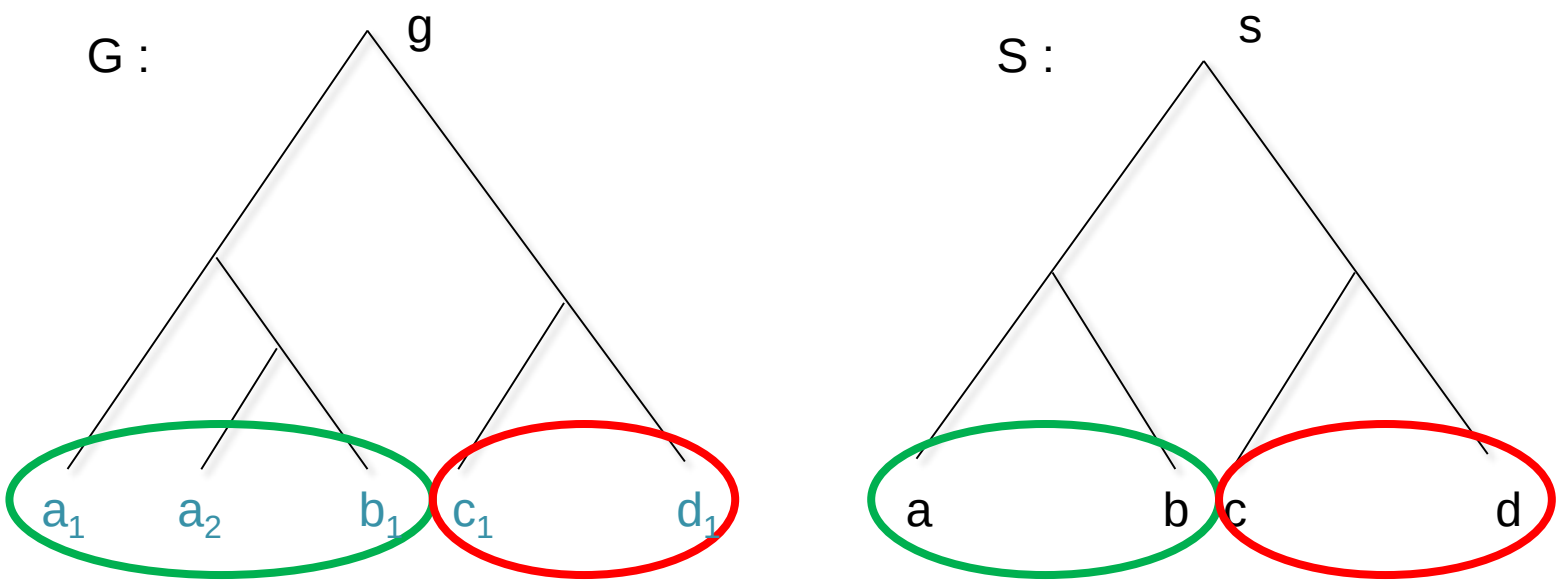
Introduction

- Speciations and duplications are typically inferred by **reconciling** G with its corresponding **species tree** S
  - Idea : map each modern gene to the species containing it, and add duplications to make G "agree" with S



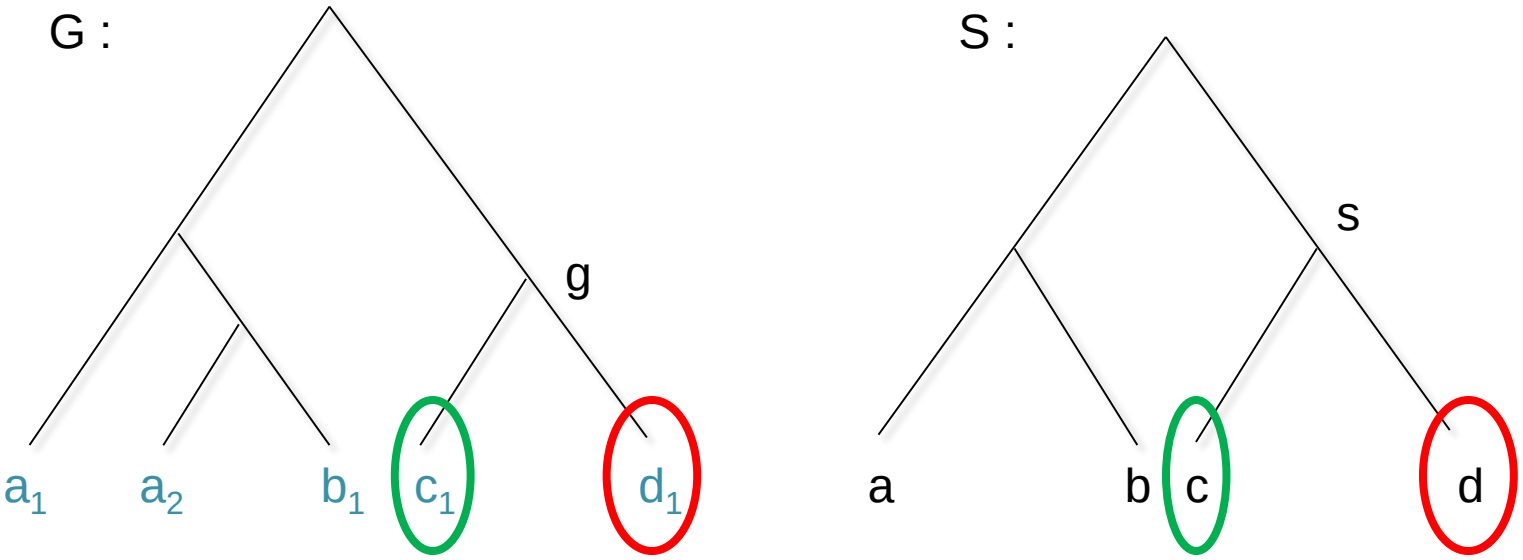
Introduction

- An internal node g of V(G) is a speciation when there is a s in V(S) such that
  - The leaves in the left subtree of g all map to leaves in the left subtree of s
  - Idem for the right side



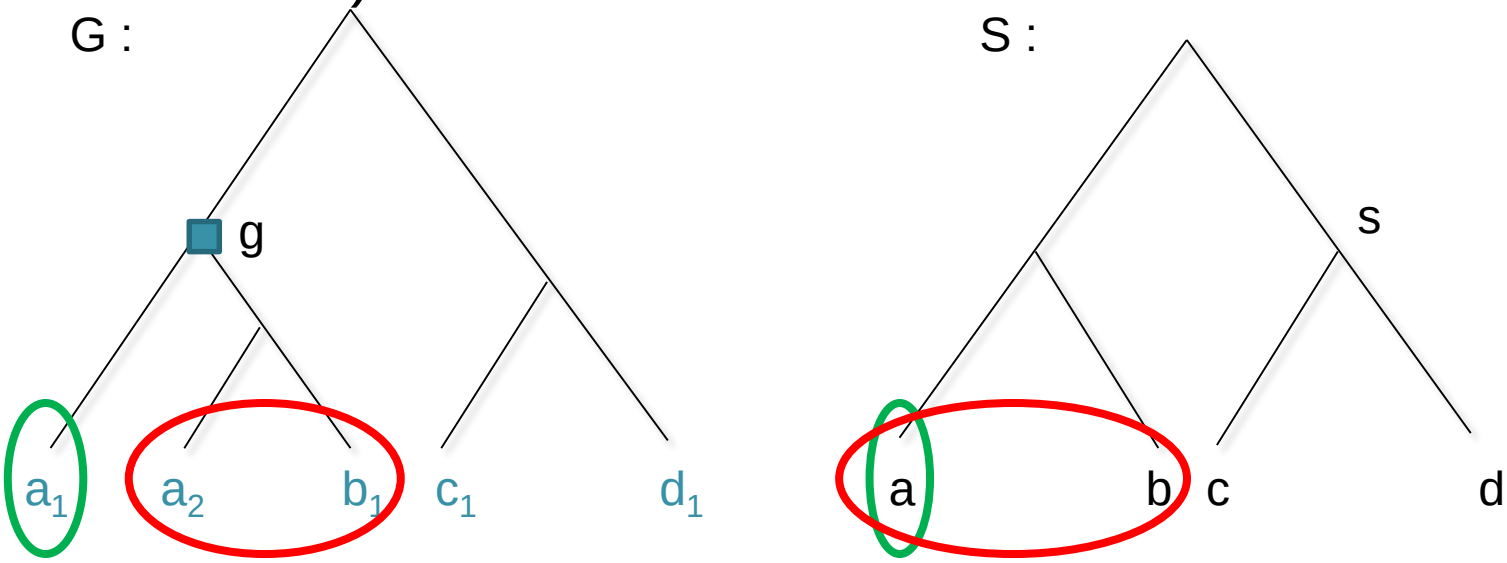
Introduction

- An internal node g of V(G) is a speciation when there is a s in V(S) such that
  - The leaves in the left subtree of g all map to leaves in the left subtree of s
  - Idem for the right side



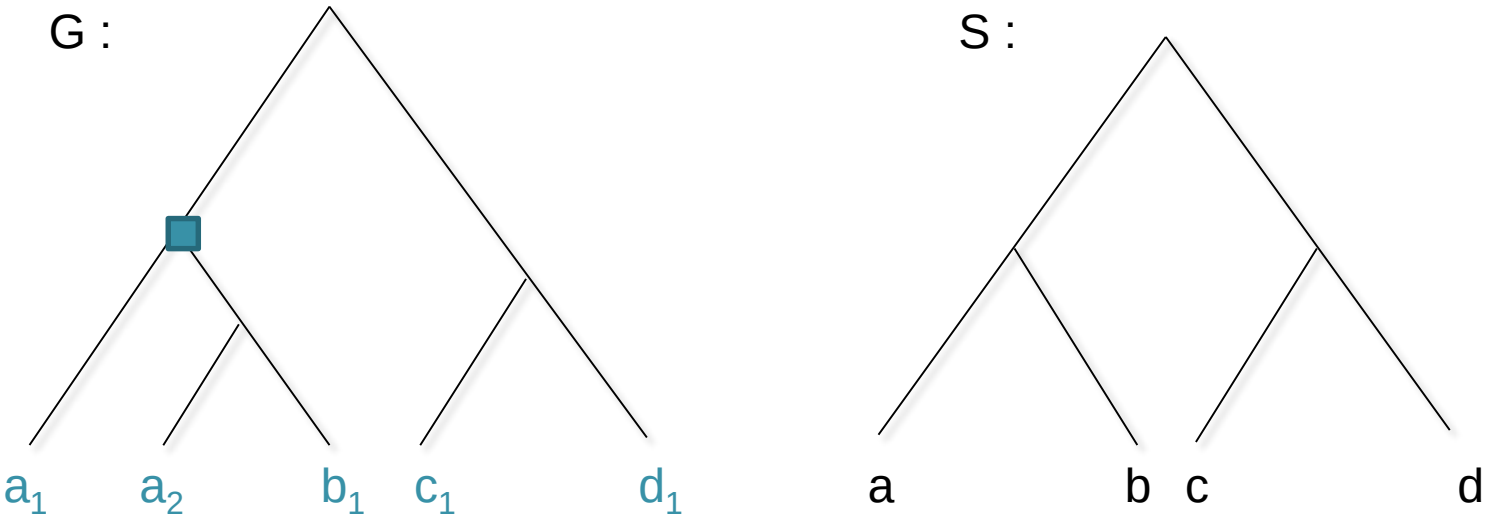
Introduction

- Otherwise, g is a duplication
  - In this case, duplication is apparent :
    - Two copies of the same gene ended up in the ‘a’ species
    - Non-apparent duplications are possible (we will se later)



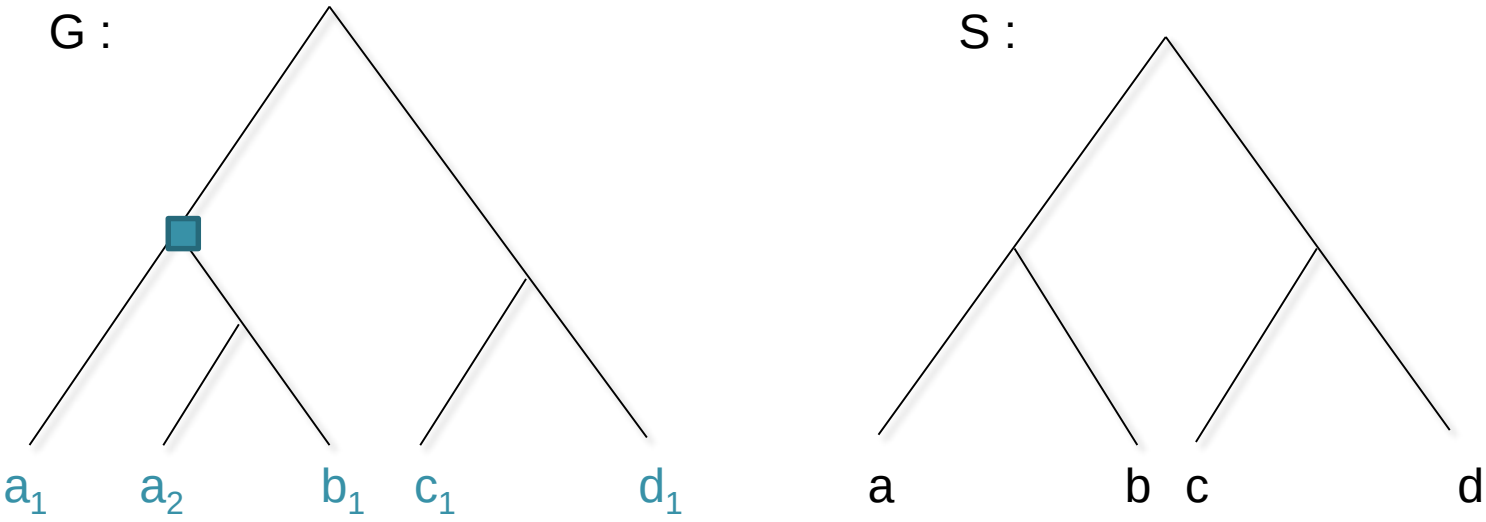
Introduction

- Suppose we are given the orthology/paralogy relationships
  - For instance, some deity lets us know that a_1, b_1 are orthologous
  - Then this gene tree is wrong !



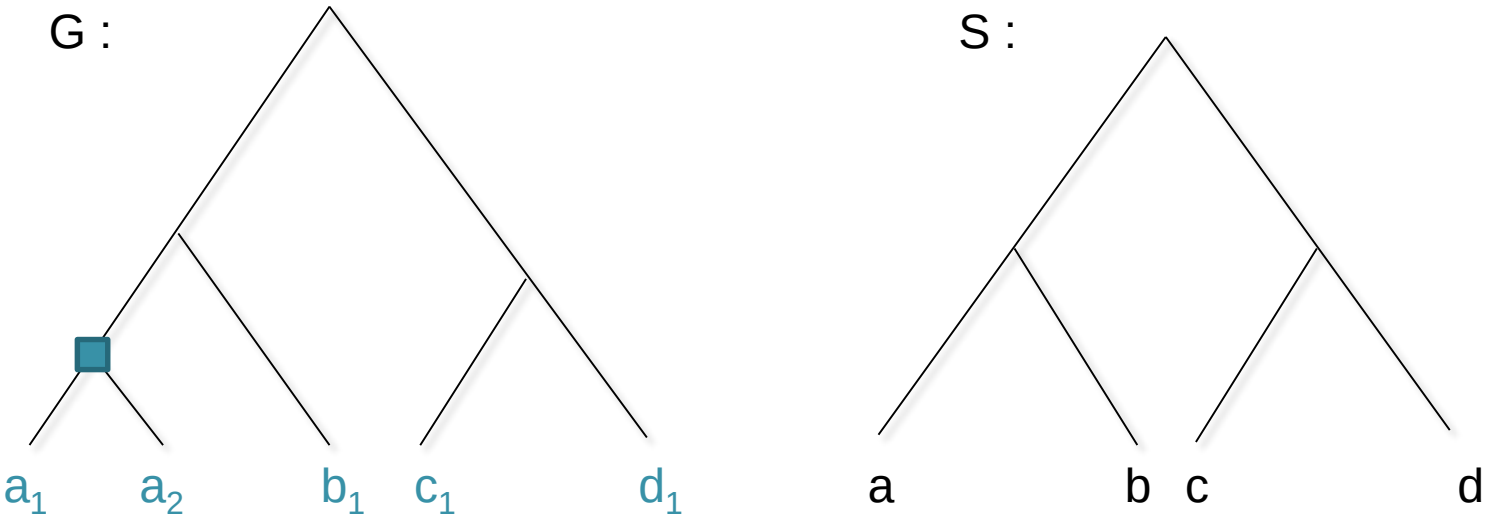
Introduction

- How can we make a_1, b_1 orthologous ?



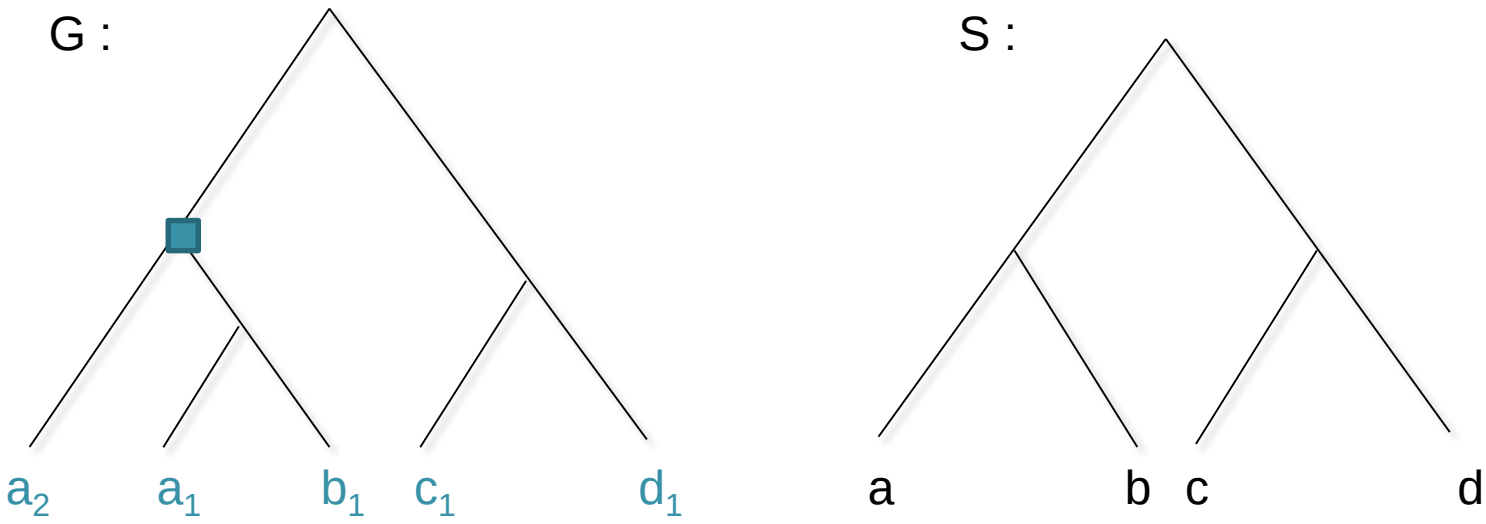
Introduction

- How can we make a_1, b_1 orthologous ?



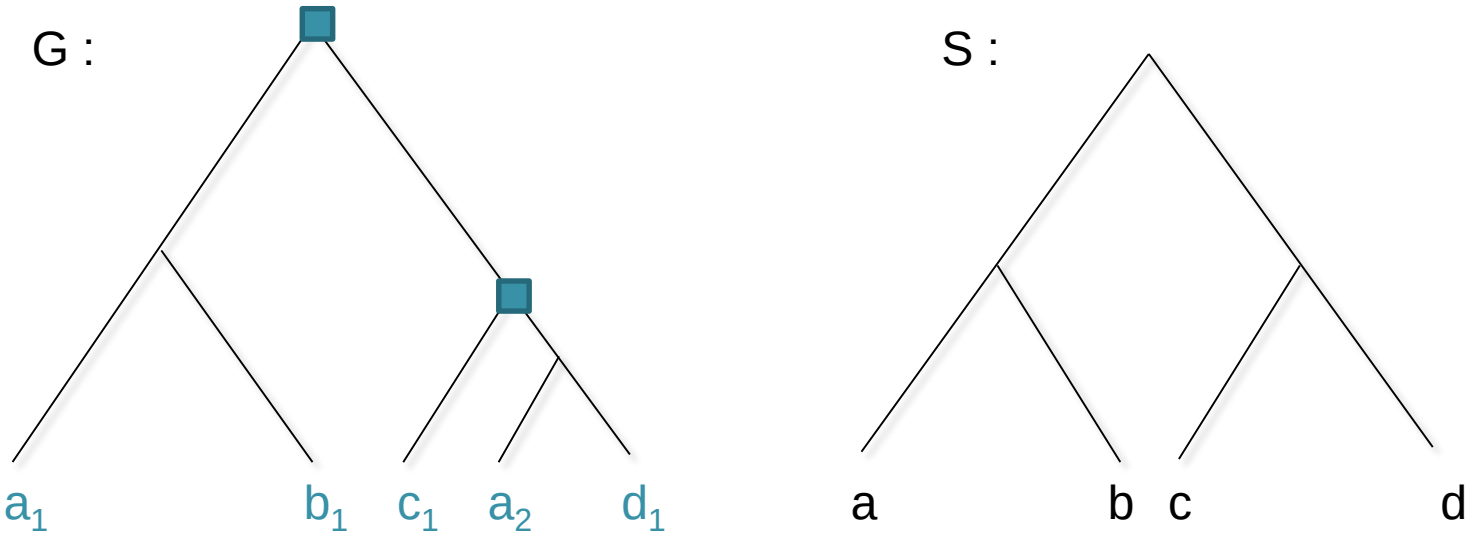
Introduction

- How can we make a_1, b_1 orthologous ?



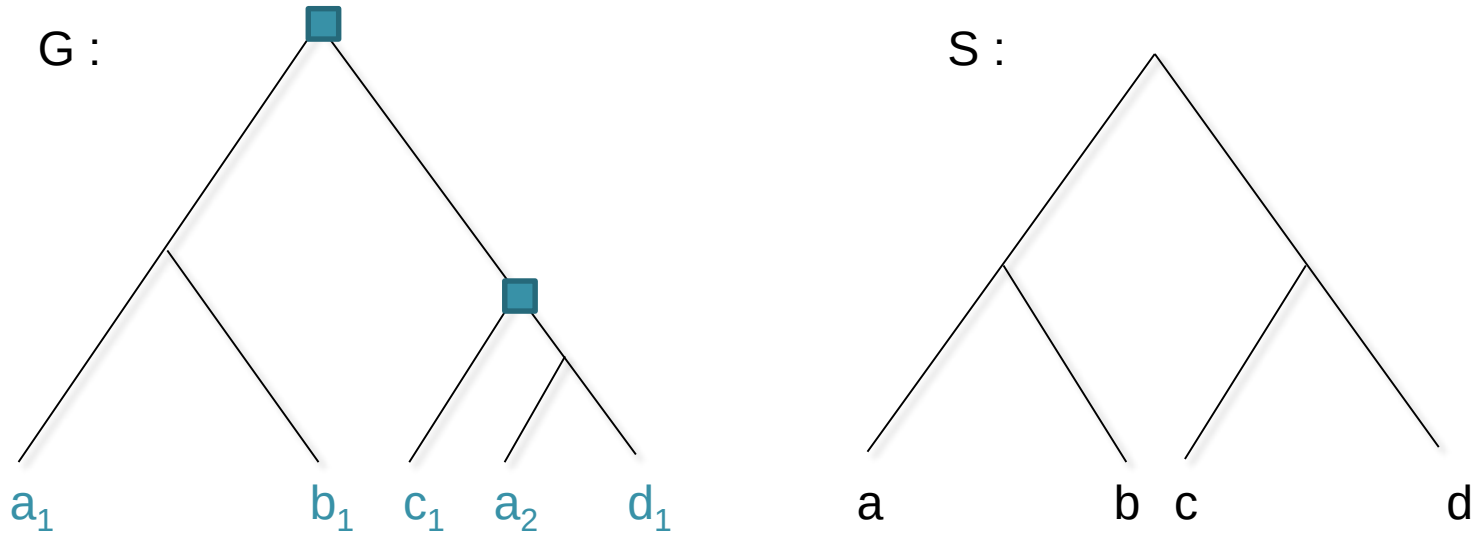
Introduction

- How can we make a_1, b_1 orthologous ?



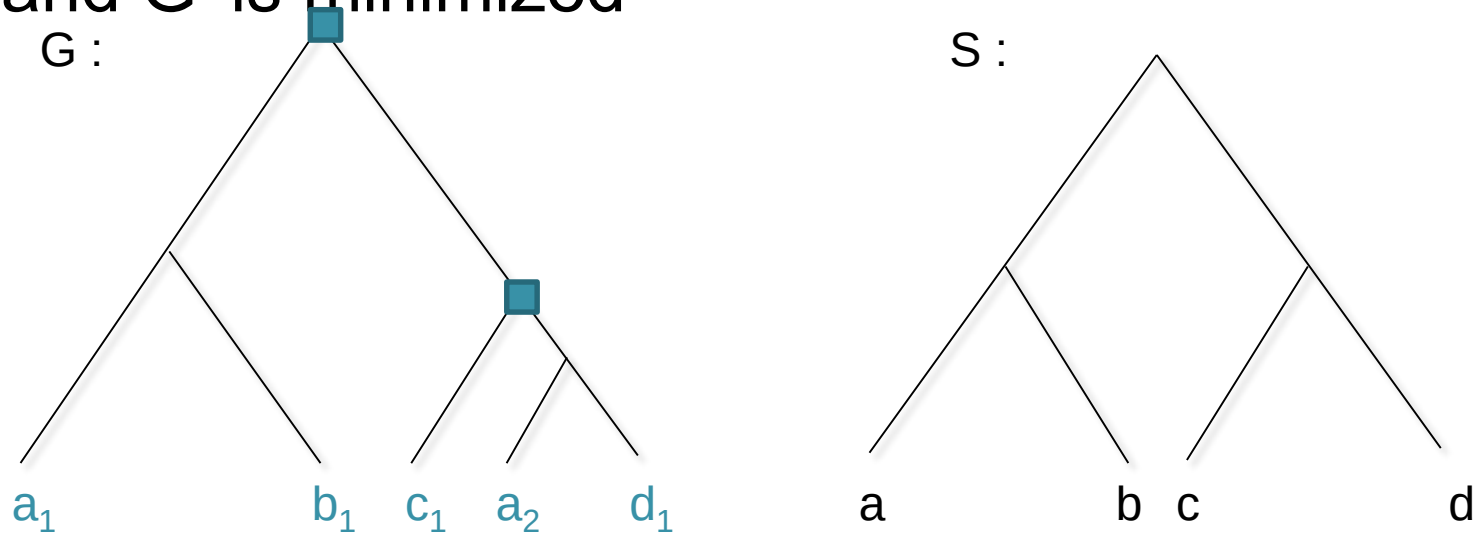
Introduction

- How can we make a_1, b_1 orthologous ?
- And mess up G as least as possible ?
- What if we're given many orthology constraints ?



Problem statement

- Given : a gene tree G, a species tree S, and a set P of pairs of genes that are required to be orthologous
- Find : a corrected gene tree G’ in which every pair (g1, g2) in P are orthologous in G’, such that the Robinson-Foulds distance between G and G’ is minimized

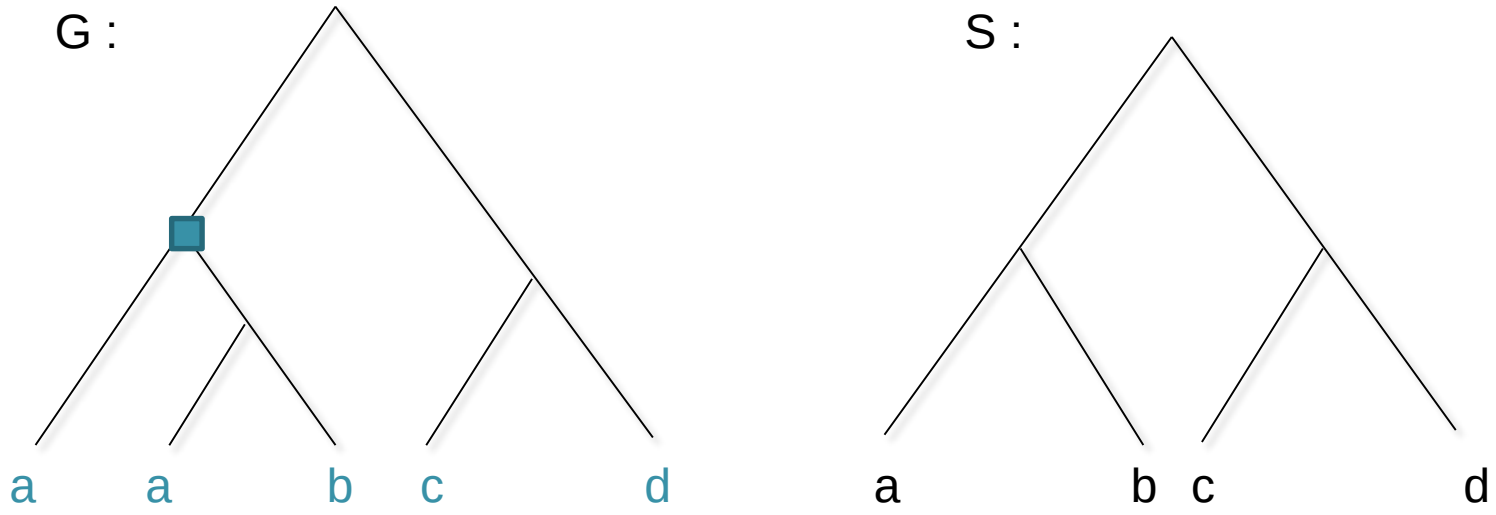


Introduction

- Two copies of the same gene were found twice in the same species (g_1, g_2)

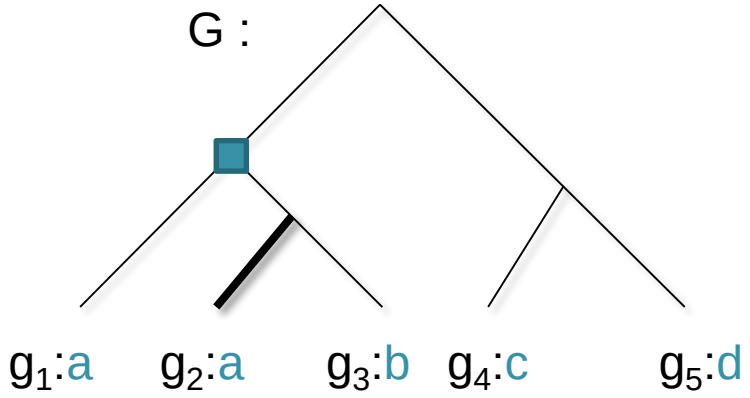
  =>

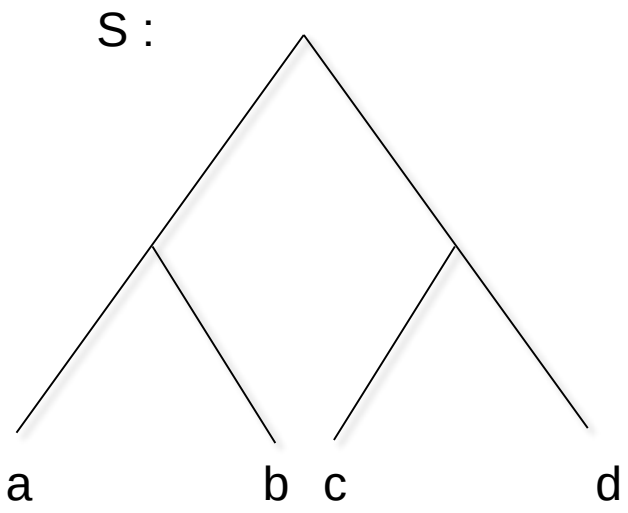
  We need to infer a duplication



Accuracy of gene trees

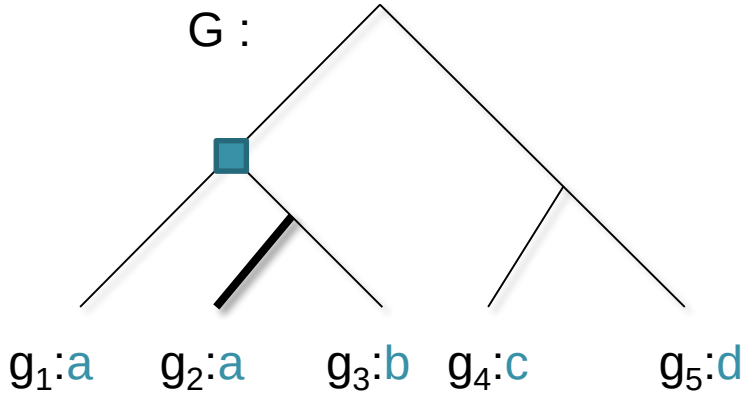
- A few misplaced leaves in G can lead to a completely different reconciliation

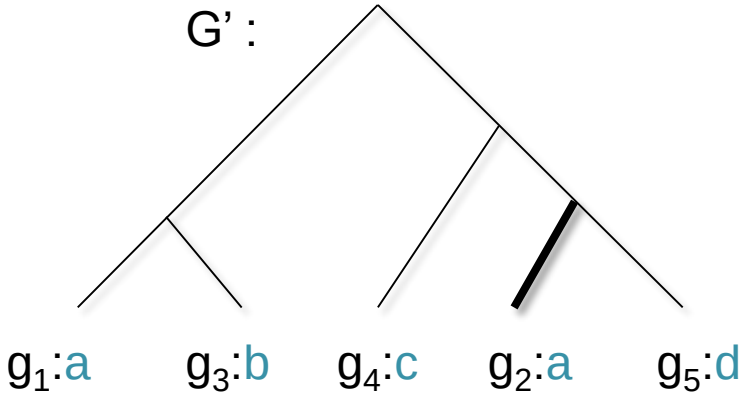


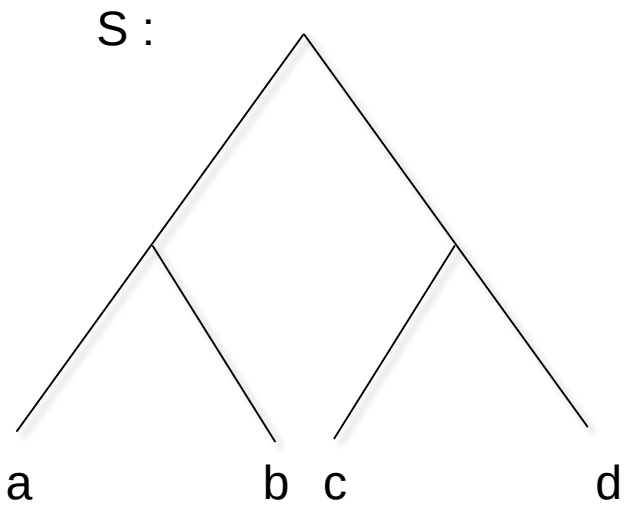


Accuracy of gene trees

- A few misplaced leaves in G can lead to a completely different reconciliation

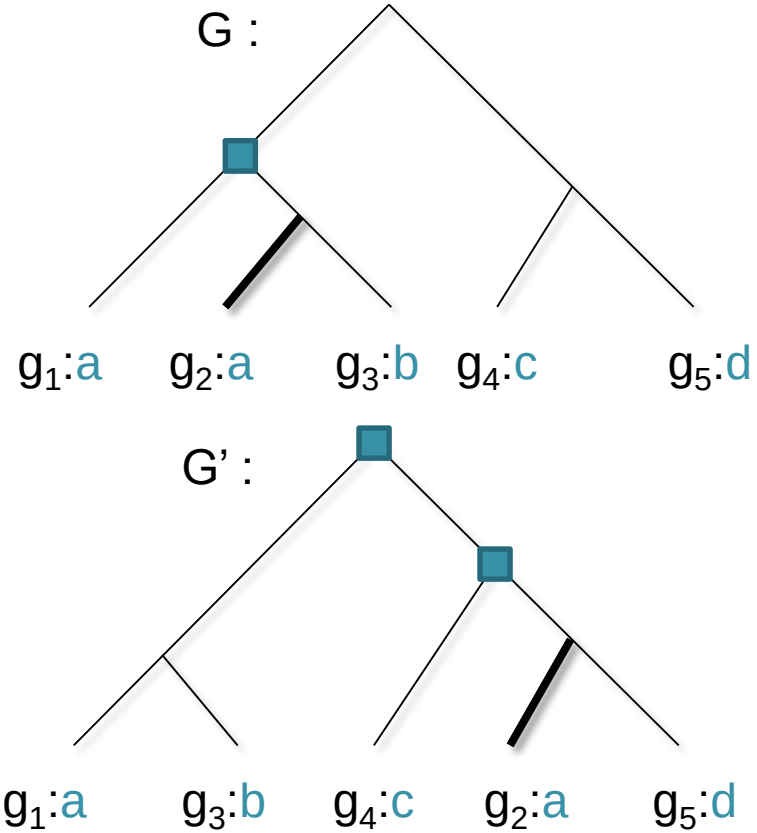


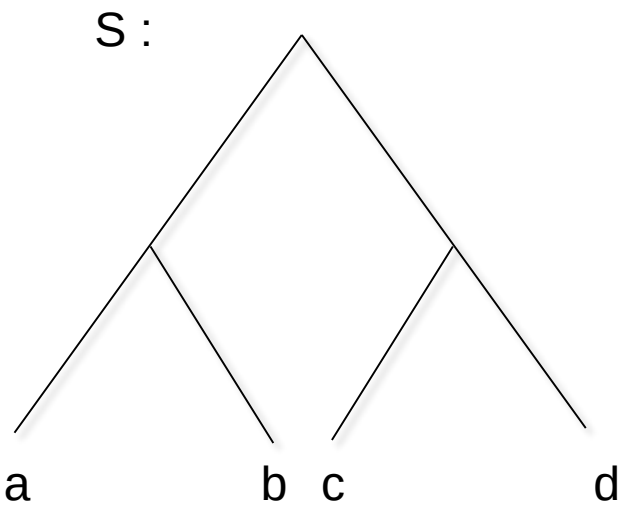




Accuracy of gene trees

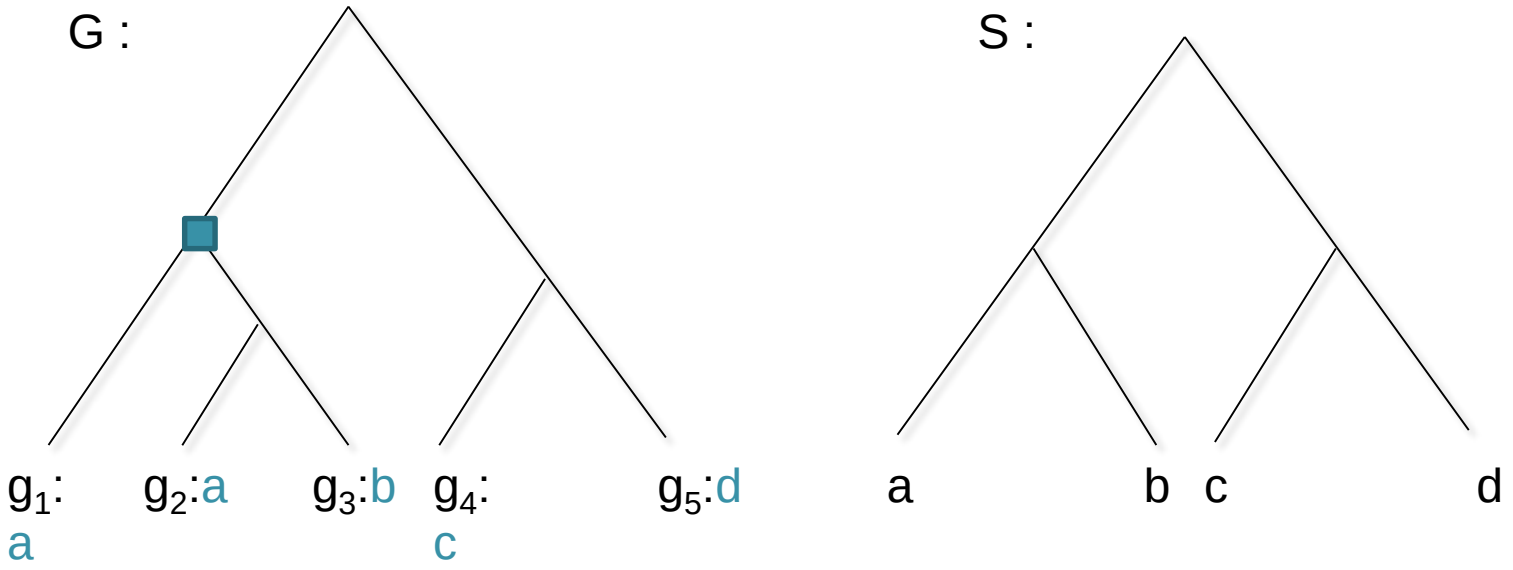
- A few misplaced leaves in G can lead to a completely different reconciliation





Accuracy of gene trees

- Inaccuracies in gene trees lead to
  - **Erroneous topologies**
  - **Erroneous orthology/paralogy relationships**
- We use **gene order** to detect and correct such errors



Gene tree inference and correction

- Some available information to infer and correct gene trees
  - Sequences (MP, ML, Bayesian, …)
  - Species tree topology (GIGA)
  - Branch/clade support (LSM)
  - Speciation/duplication events inferred by reconciliation (TreeBeST)
  - Gene synteny (SYNERGY)
  - **Gene position and order on genome**

Gene order

- **Genome** : a string of genes, giving the order in which genes are found in a given species
  - Genome for X species : “a b c d e f g …”
- **Region** : a subsequence of a genome
  - Pick a subset of a genome’s genes, maintaining the order
  - a **b c** d **e** f **g** h ...

    =>

    b c e g region
- Typically, we impose a limit on the size of a region and on the genome distance between its members

Region homology

- Two **genes are homologous** if they descend from a common ancestral gene
  - This ancestral has undergone speciation or duplication

Region homology

- Two **genes are homologous** if they descend from a common ancestral gene
  - This ancestral has undergone speciation or duplication
- Can we define **region homology** similarly?

Region homology

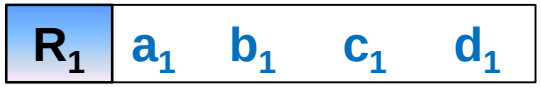
- Two **genes are homologous** if they descend from a common ancestral gene, which has undergone speciation or duplication
- Can we define **region homology** similarly ?
- Two **regions are homologous** if they descend from a **common ancestral region**, which has undergone **speciation** or **duplication**

Region homology

- Two **genes are homologous** if they descend from a common ancestral gene, which has undergone speciation or duplication
- Can we define **region homology** similarly ?
- Two **regions are homologous** if they descend from a **common ancestral region**, which has undergone **speciation** or **duplication**
  - What does that even mean ?

Region homology

- Common ancestral region
  - For two given regions R_1, R_2

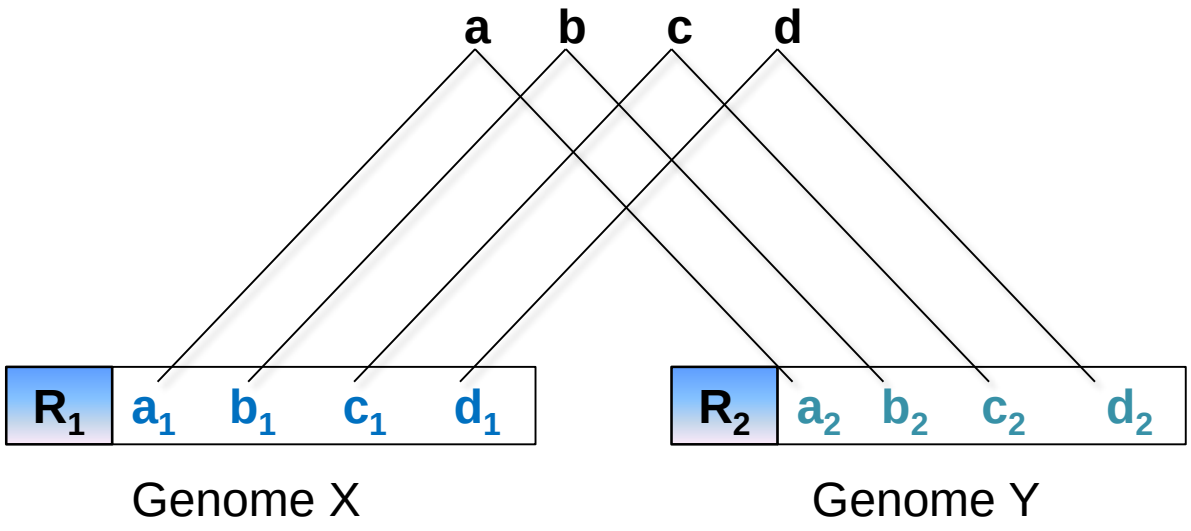


Genome X

Genome Y

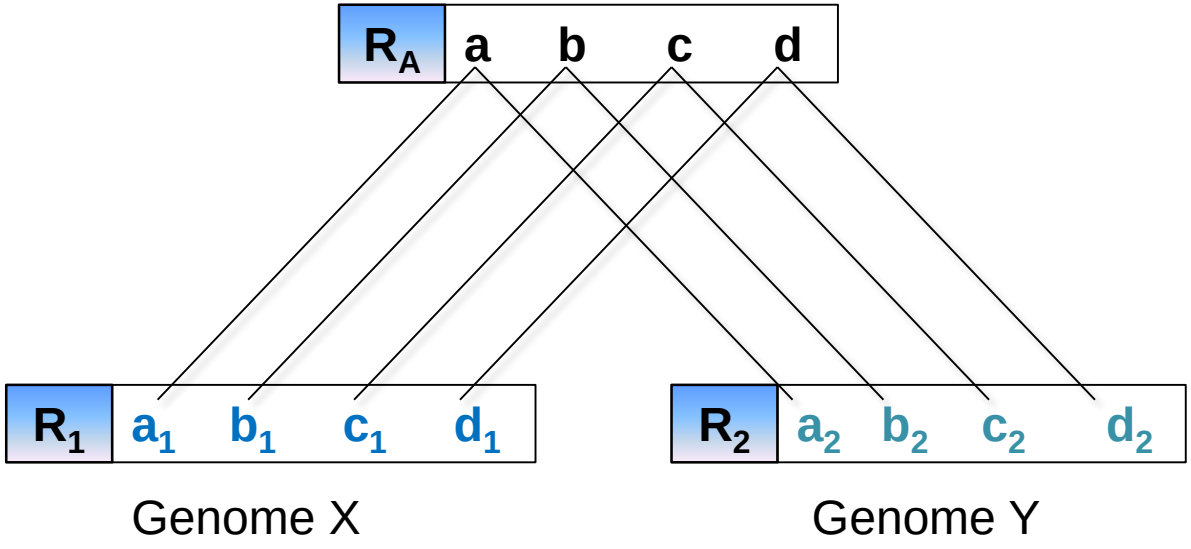
Region homology

- Common ancestral region
  - For two given regions R_1, R_2
    - Subdivide their genes into gene families F_1, F_2, …, F_n
      - In the example, four families (a,b,c,d)
    - Look at the roots of the gene trees for all the F_i's



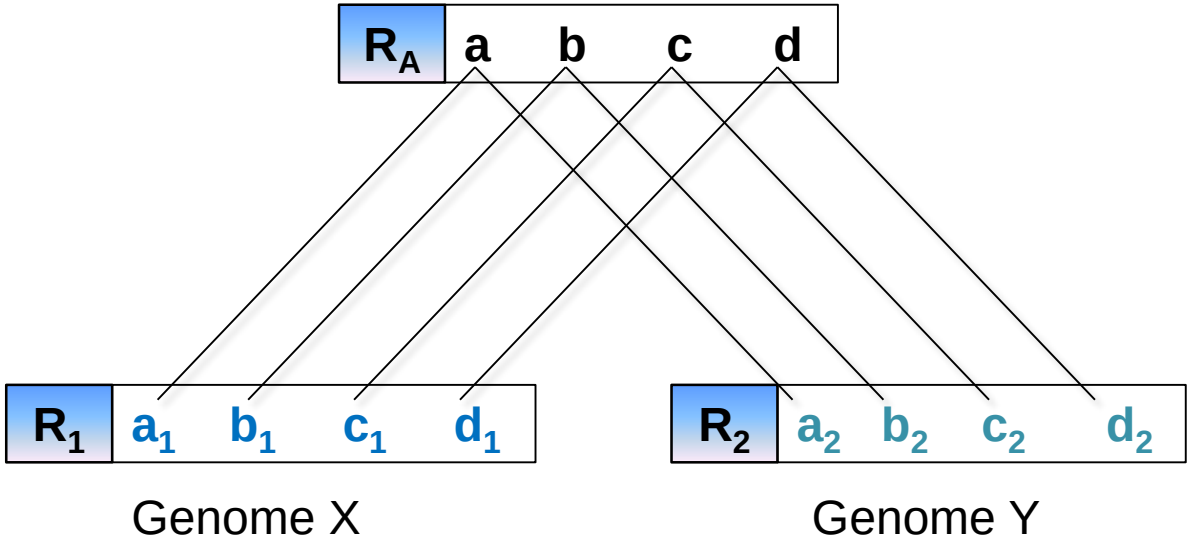
Region homology

- Common ancestral region
  - If all these ancestral genes are in the same ancestral genome, R_1, R_2 share a common ancestral region R_A



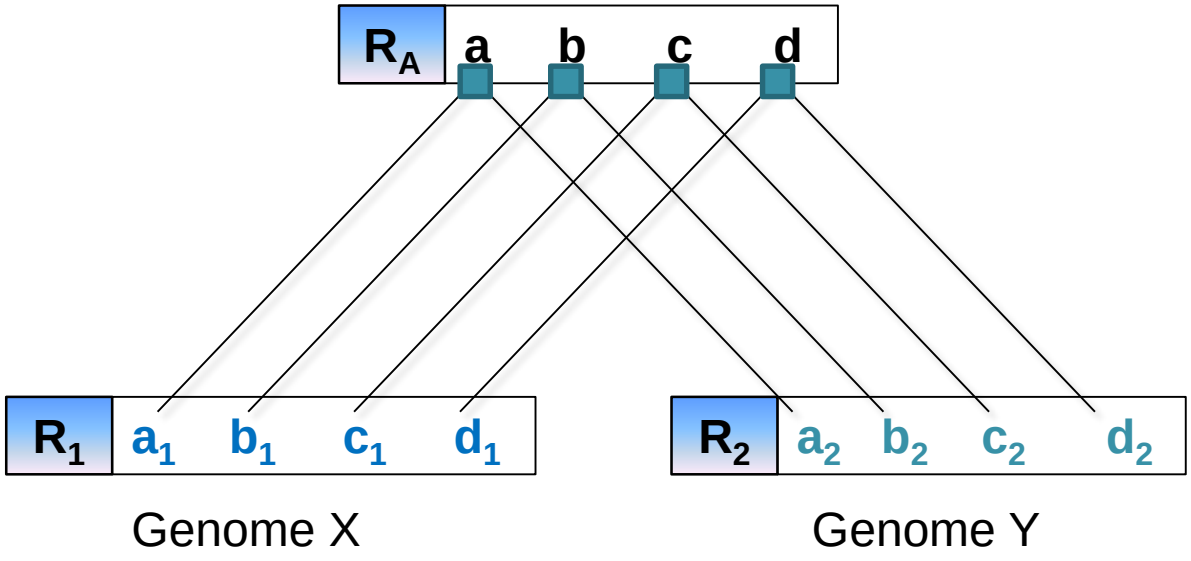
Region homology

- Region speciation
  - All the roots are speciation



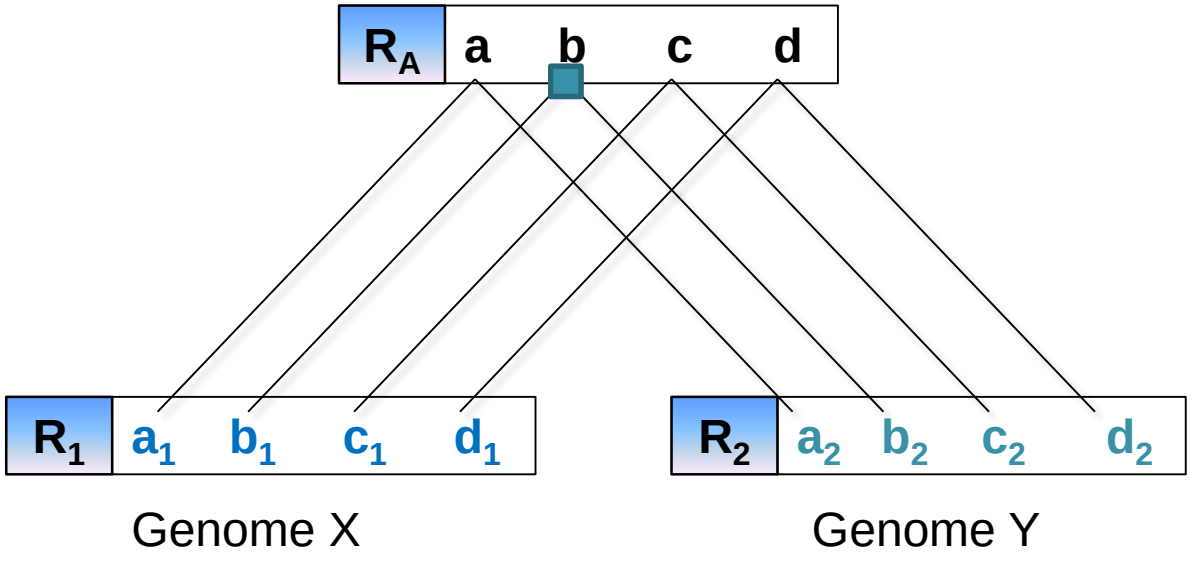
Region homology

- Region duplication
  - All the roots are duplications
  - Corresponds to a segmental duplication (or “region duplication” in the ancestral genome



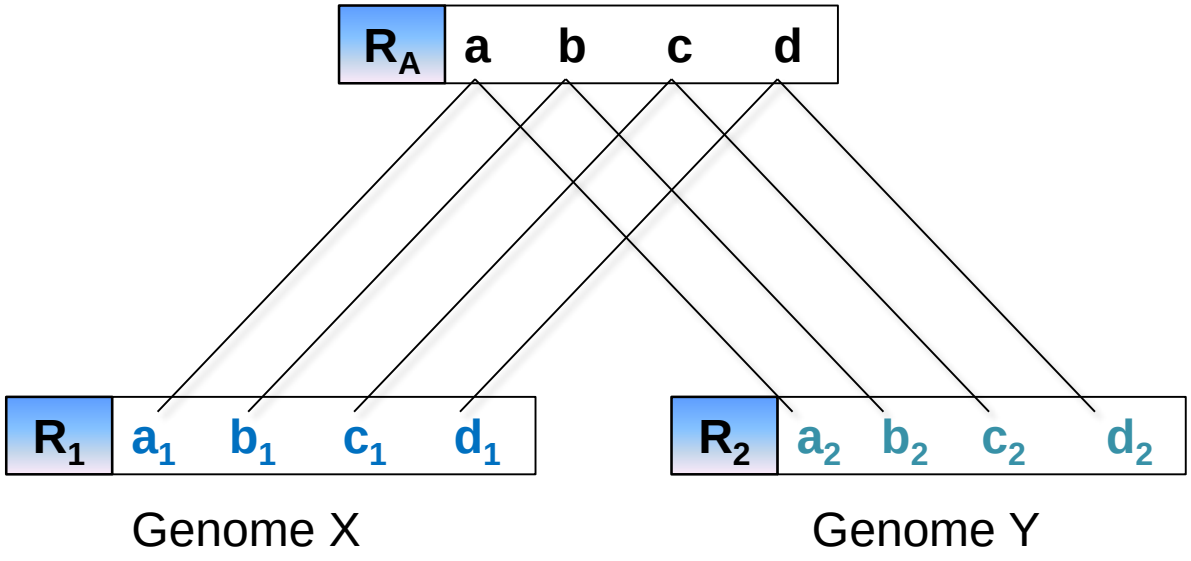
Region homology

- Not homologous regions



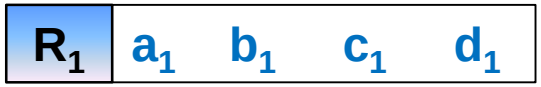
No convergent evolution hypothesis

- Hypothesis : similar regions are homologous



Homology contradiction

- If we find two similar regions and look at the roots of the gene family trees, we expect them all to be the same type

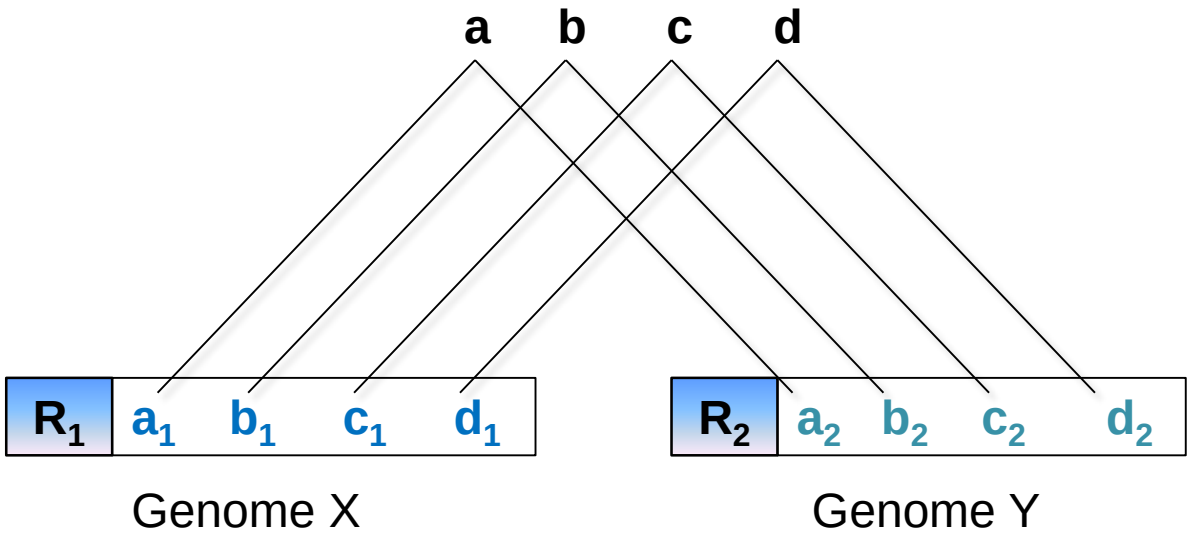


Genome X

Genome Y

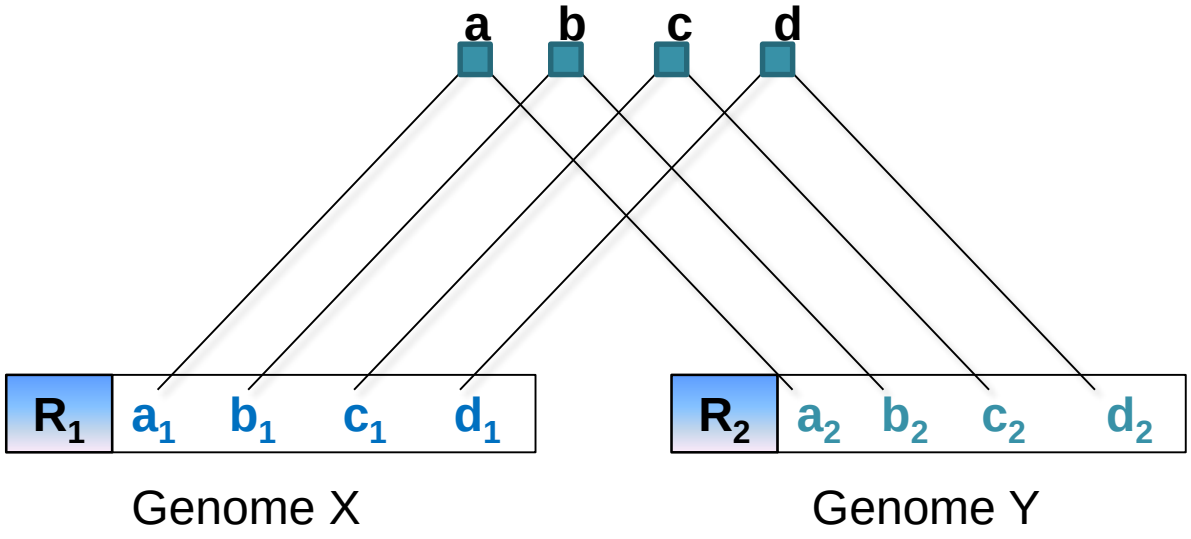
Homology contradiction

- If we find two similar regions and look at the roots of the gene family trees, we expect them all to be the same type



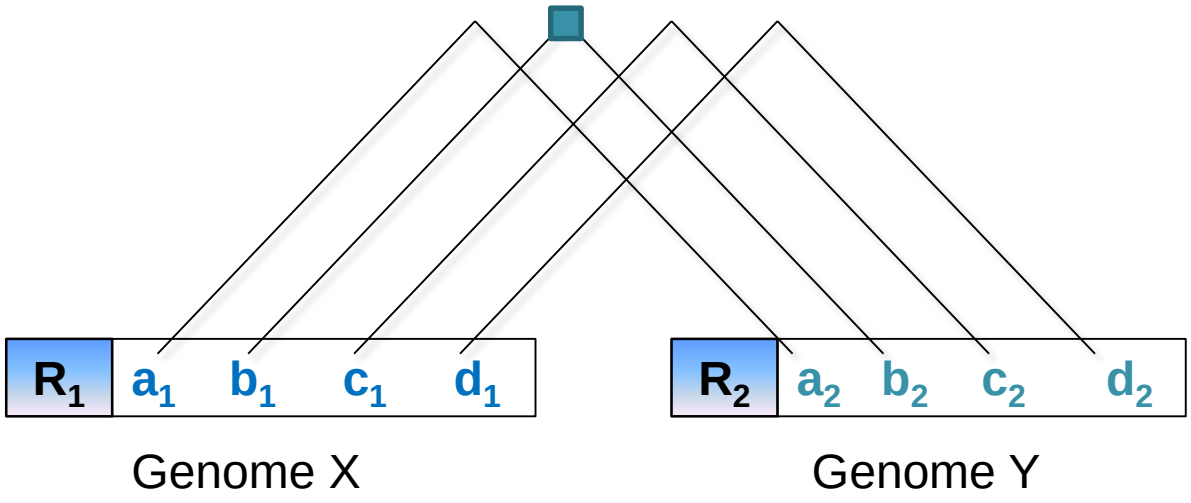
Homology contradiction

- If we find two similar regions and look at the roots of the gene family trees, we expect them all to be the same type



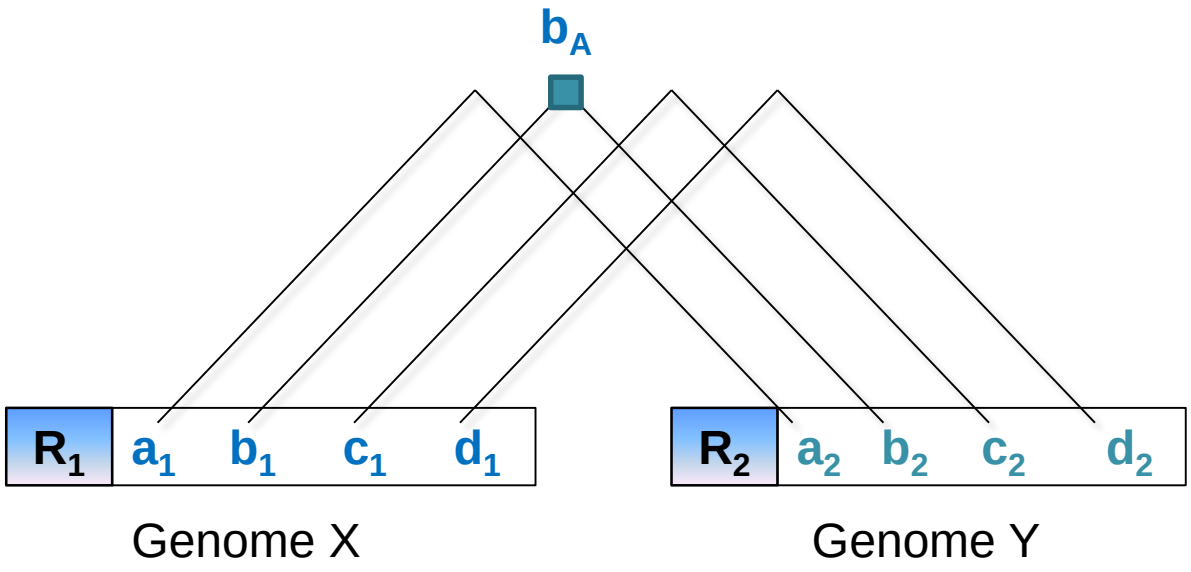
Homology contradiction

- Otherwise, there is a homology contradiction (an error in one of the gene trees)



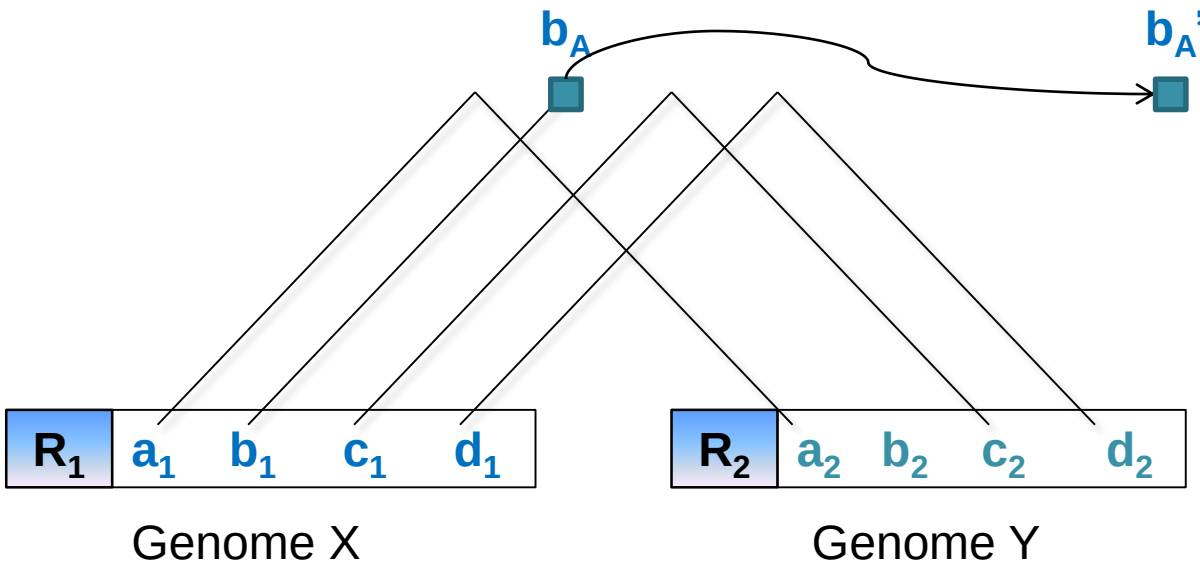
Homology contradiction

- Why not ?
  - If b_A duplicated, the copy typically went somewhere else on the ancestral genome



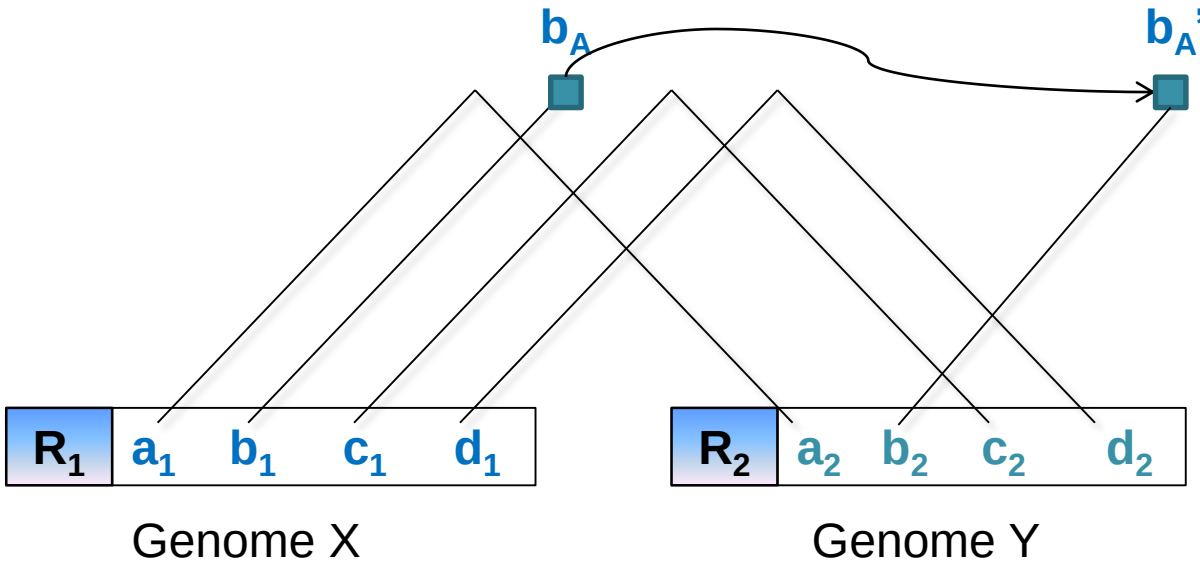
Homology contradiction

- Why not ?
  - If b_A duplicated, the copy typically went somewhere else on the ancestral genome



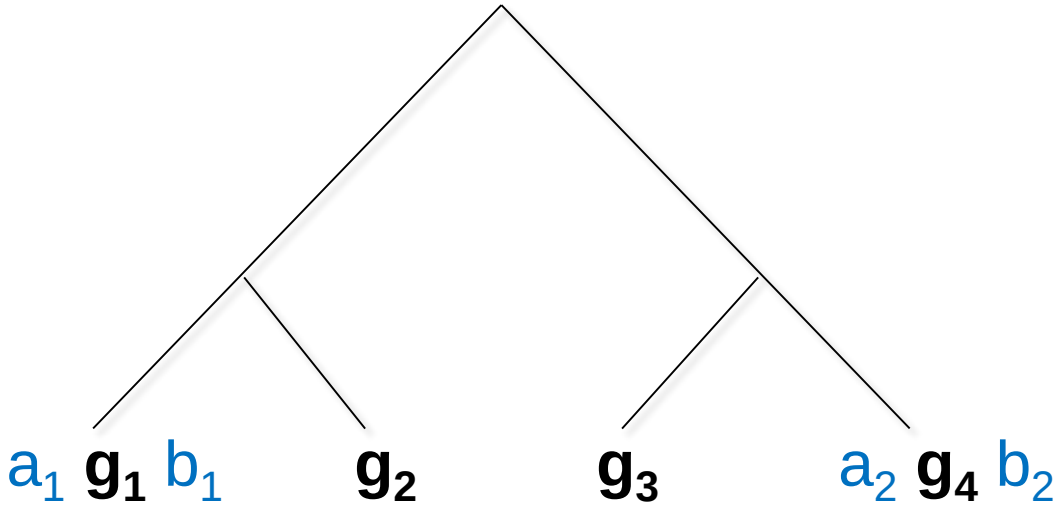
Homology contradiction

- Why not ?
  - If b_A duplicated, the copy typically went somewhere else on the ancestral genome
  - And somehow, during evolution, it ended up in a region similar to R_1, mostly by **chance**



Strong no convergent evolution

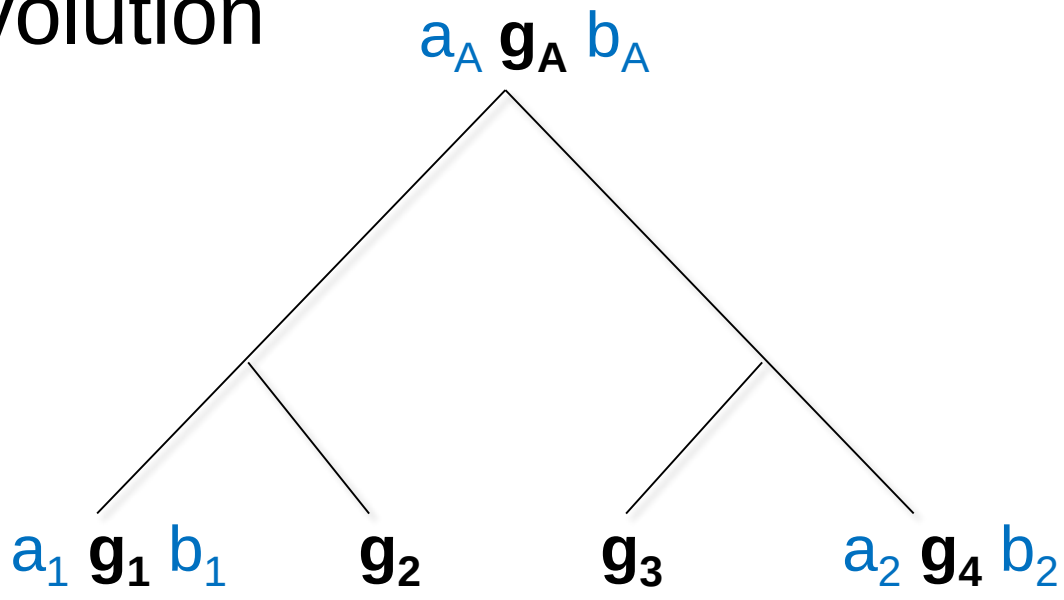
- Hypothesis : similarity is inherited from the common ancestral region, and is preserved during the course of evolution



G : gene tree for g family

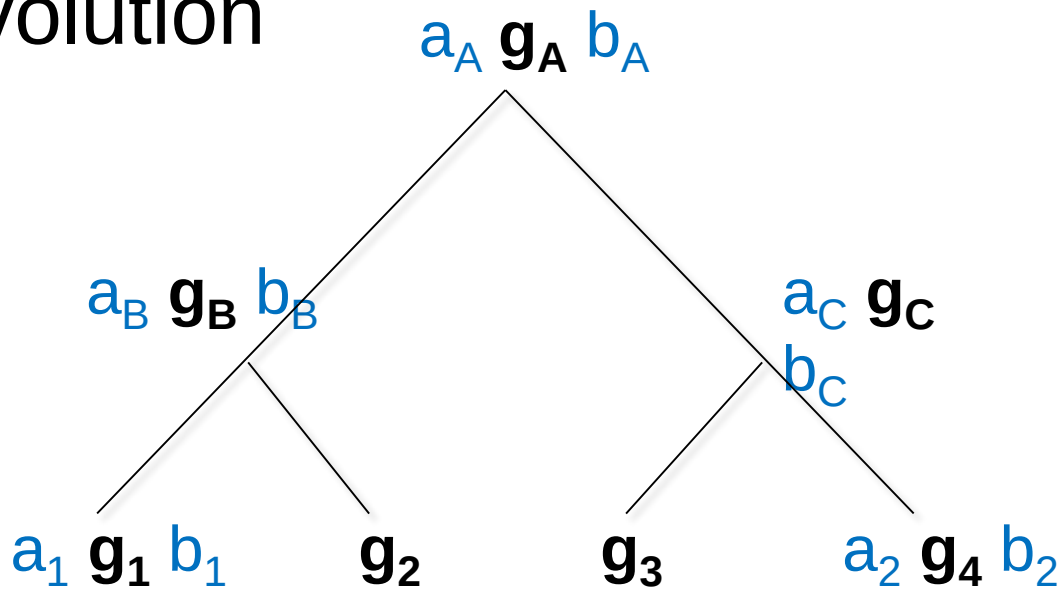
Strong no convergent evolution

- Hypothesis : similarity is inherited from the common ancestral region, and is preserved during the course of evolution



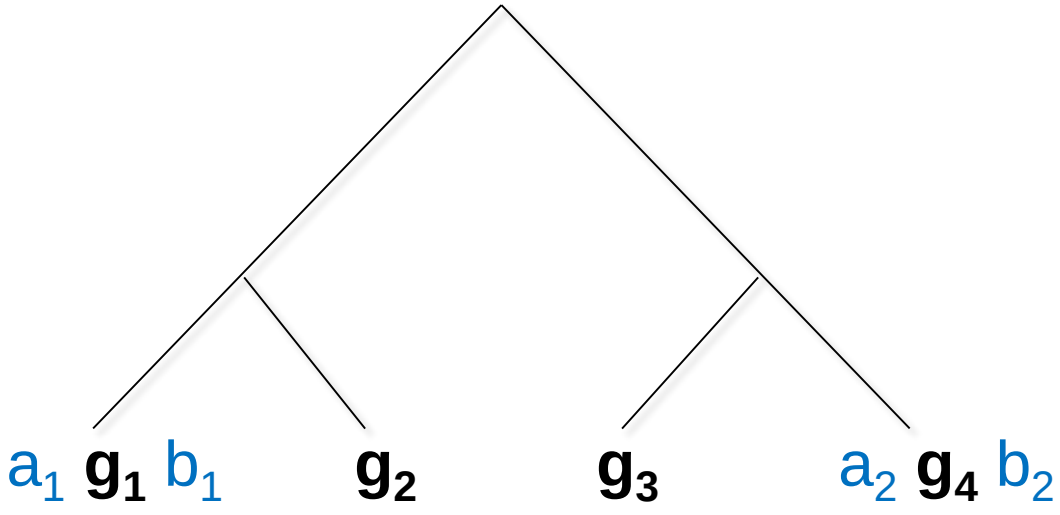
Strong no convergent evolution

- Hypothesis : similarity is inherited from the common ancestral region, and is preserved during the course of evolution



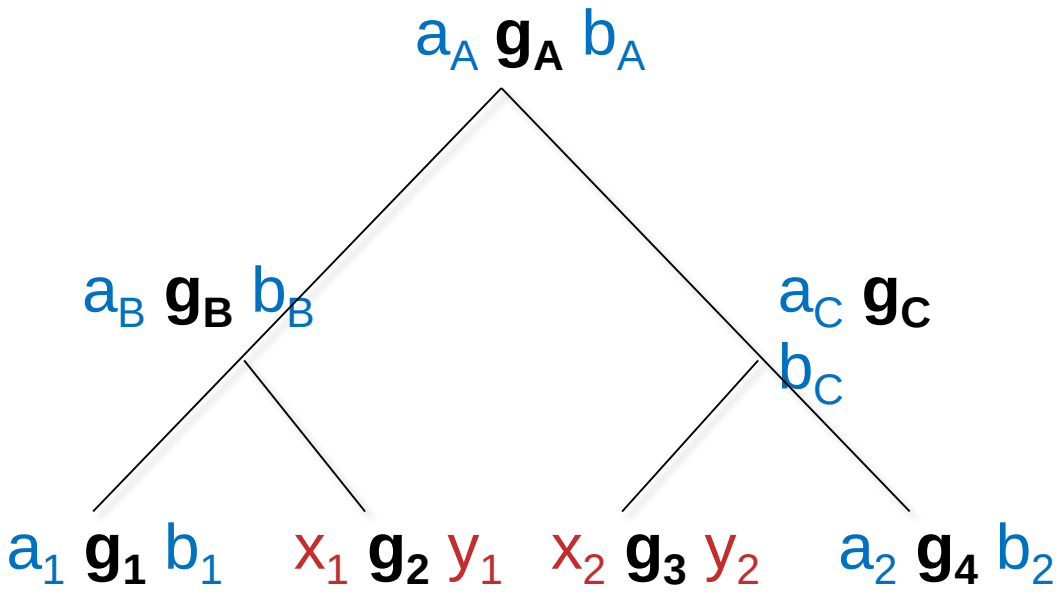
Strong no convergent evolution

- Otherwise, we must assume g_1 and g_2 gained their region similarity by **chance**



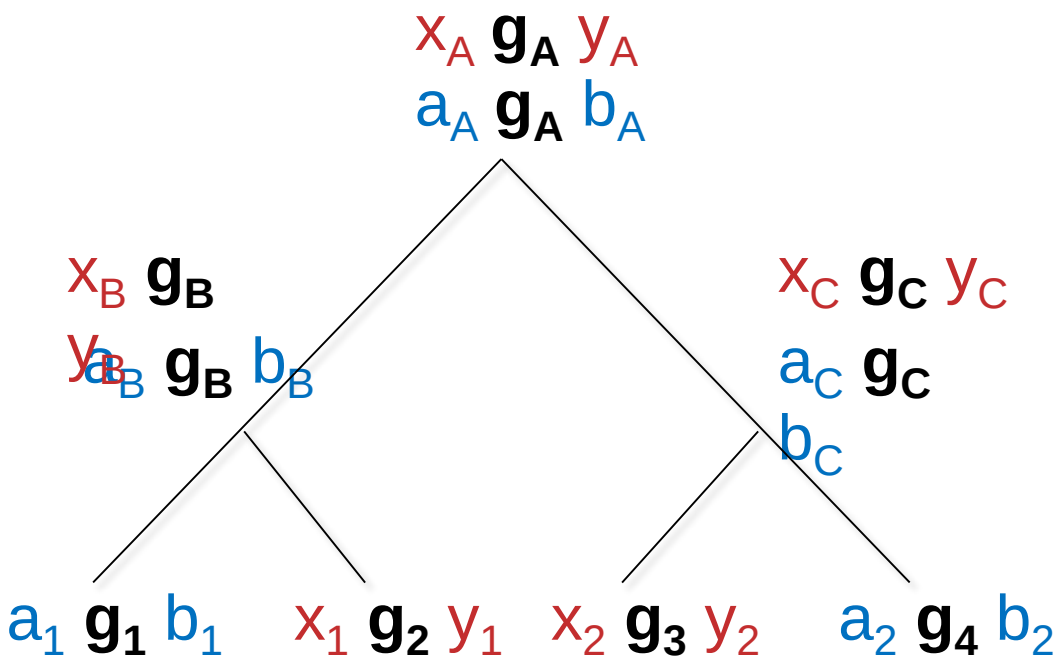
Region overlapping

- Two ancestral genes may belong to two different region families simultaneously



Region overlapping

- Two ancestral genes may belong to two different region families simultaneously



Results

- We looked for homology contradictions and context overlapping in ~6000 Ensembl gene trees
- All trees contained genes for the Zebrafish, Medaka, Stickleback and Tetraodon species, and we included Human and Mouse as outgroups

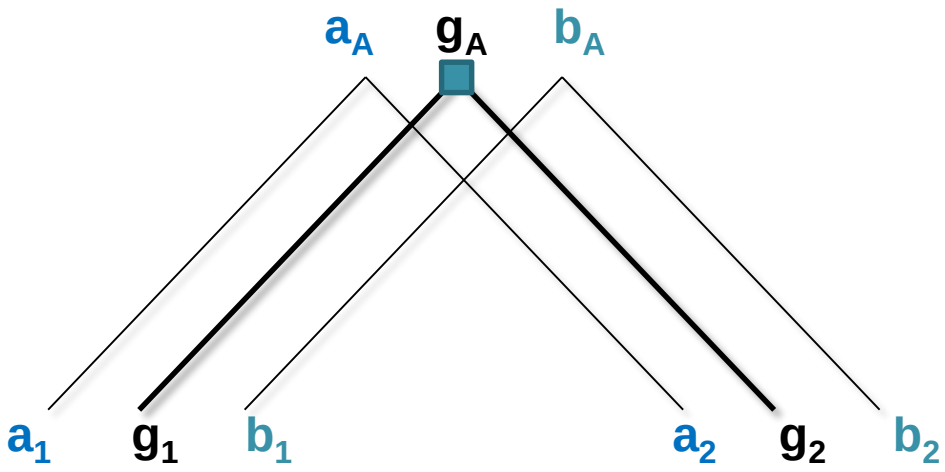
Results

- Each gene was assigned a size 3 region
  - Triplet containing the gene, and its **left/right adjacencies**
- The central gene is the gene of interest

- Two regions ($a\ g_1\ b$), ($x\ g_2\ y$) are homologous if a, x are in the same family, as well as b, y

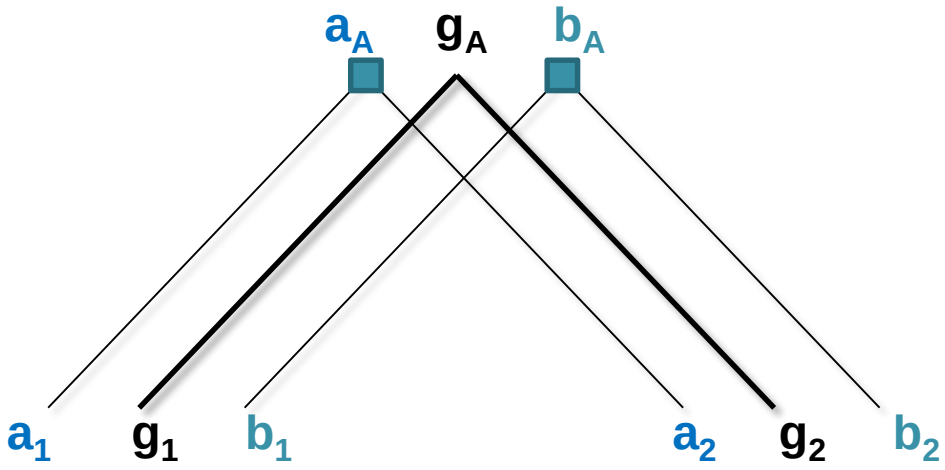
Results

- Paralogy contradiction
  - g_A should not be a duplication



Results

- Orthology contradiction
  - g_A should not be a speciation



Results

Number of trees	6241
Paralogy contradiction	22.5 % (1407 trees)
Orthology contradiction	10.8 % (677 trees)
Region overlap	3.4 % (210 trees)
At least one contradiction	31.3 % (1959 trees)

Table 1 : Number of Ensembl gene trees with errors

Results

Number of trees	6241
Paralogy contradiction	22.5 % (1407 trees)
Orthology contradiction	10.8 % (677 trees)
Region overlap	3.4 % (210 trees)
At least one contradiction	31.3 % (1959 trees)

Table 1 : Number of Ensembl gene trees with errors

77% of paralogy contradictions correspond to duplications marked as "**dubious**" by Ensembl

(dubious are Non-Apparent Duplications)

Gene tree correction

- How should such errors be corrected ?
- We need to find an **error-free** gene tree with **equal or better** statistical support
  - Explore the original gene tree's neighborhood
  - Algorithmically free the gene tree from errors, minimizing some criteria

Gene tree correction

- How should such errors be corrected ?
- We need to find an **error-free** gene tree with **equal or better** statistical support
  - Explore the original gene tree's neighborhood
  - Algorithmically free the gene tree from errors, minimizing some criteria
    - Distance from original tree (NNI, SPR, TBR, RF, …)
    - Reconciliation cost
    - Get rid of dubious duplications
    - …?

Gene tree correction

- How should such errors be corrected ?
- We need to find an **error-free** gene tree with **equal or better** statistical support
  - Explore the original gene tree's neighborhood
  - Algorithmically free the gene tree from errors, minimizing some criteria
    - Distance from original tree (NNI, SPR, TBR, RF, …)
    - Reconciliation cost
    - Get rid of dubious duplications

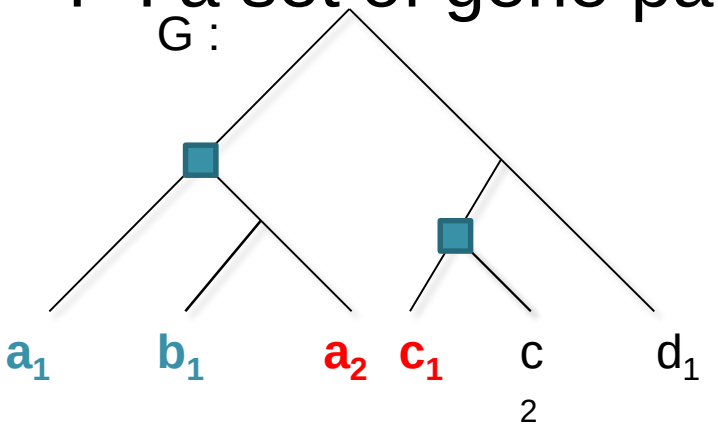
Does an error-free tree even exist ?

Gene tree correction

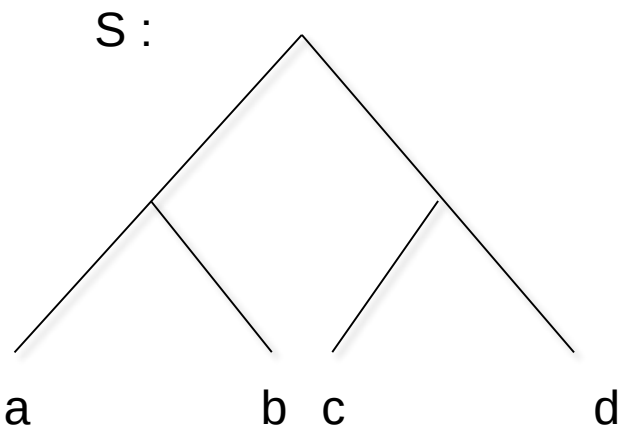
- For homology contradictions
  - R : a set of gene pairs that **must be orthologs**
  - P : a set of gene pairs **must be paralogs**

Gene tree correction

- R : a set of gene pairs that **must be orthologs**
- P : a set of gene pairs **must be paralogs**

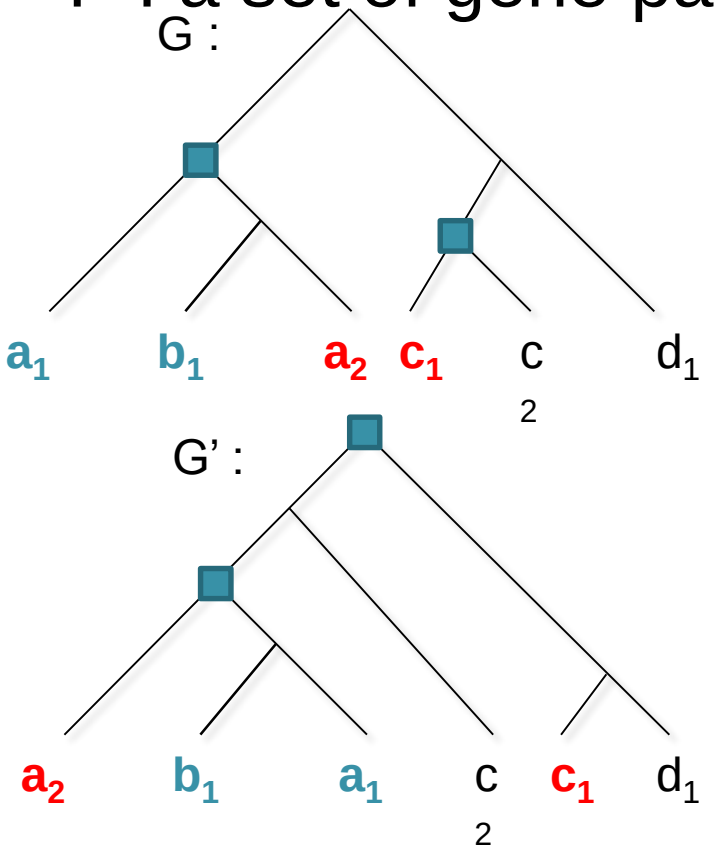


R = {(a_1, b_1)} P = {(a_2, c_1)}

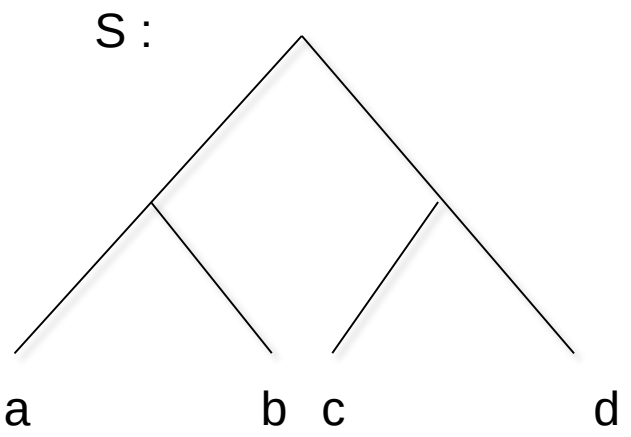


Gene tree correction

- R : a set of gene pairs that **must be orthologs**
- P : a set of gene pairs **must be paralogs**



$R = \{(a_1, b_1)\}$ $P = \{(a_2, c_1)\}$



Gene tree correction

- R : a set of gene pairs that **must be orthologs**
- P : a set of gene pairs **must be paralogs**

- It is possible to have R and P such that no gene tree can satisfy all constraints
  - Deciding if R and P are satisfiable : complexity unknown

Correction of paralogy contradictions

- Input : a gene tree G, a species tree S, and R a set of gene pairs that must be orthologs

- Output : a corrected gene tree G' in which
  - every required orthologs in R are orthologs in G'
  - **Robinson-Foulds distance** between G, G' is minimized (among all possible solutions)

- Feasible in polynomial time

Conclusion

- 3 types of errors in gene trees
  - Paralogy contradiction
  - Orthology contradiction
  - Context overlap
- How can we free a gene tree from such errors in order to get more accurate trees ?
- Do unsatisfiable constraints exist in real data ? If so, how can we interpret them ?