ORTHOLOGY RELATION AND GENE TREE CORRECTION: COMPLEXITY RESULTS

Manuel Lafond, Nadia El-Mabrouk

University of Montreal

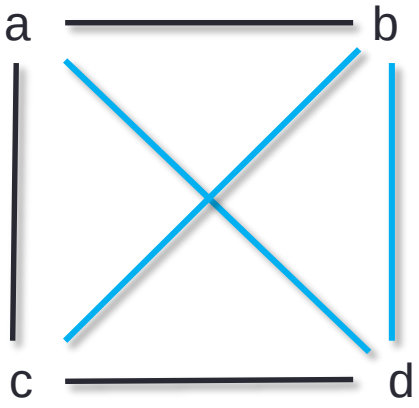
- **Two classes of problems :**
  - Orthology / paralogy relation correction
  - Gene tree correction (based on orthology / paralogy)

- **Four problems, all NP-Complete**

Orthology / paralogy correction

Makes no sense
(we'll see why)

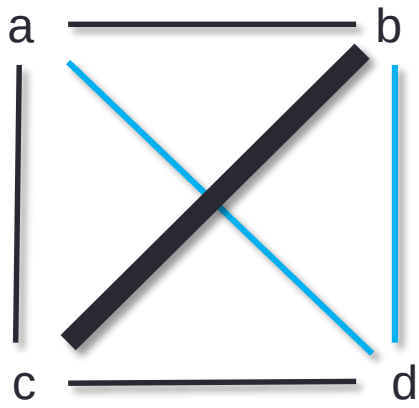
R



Correction

Makes sense

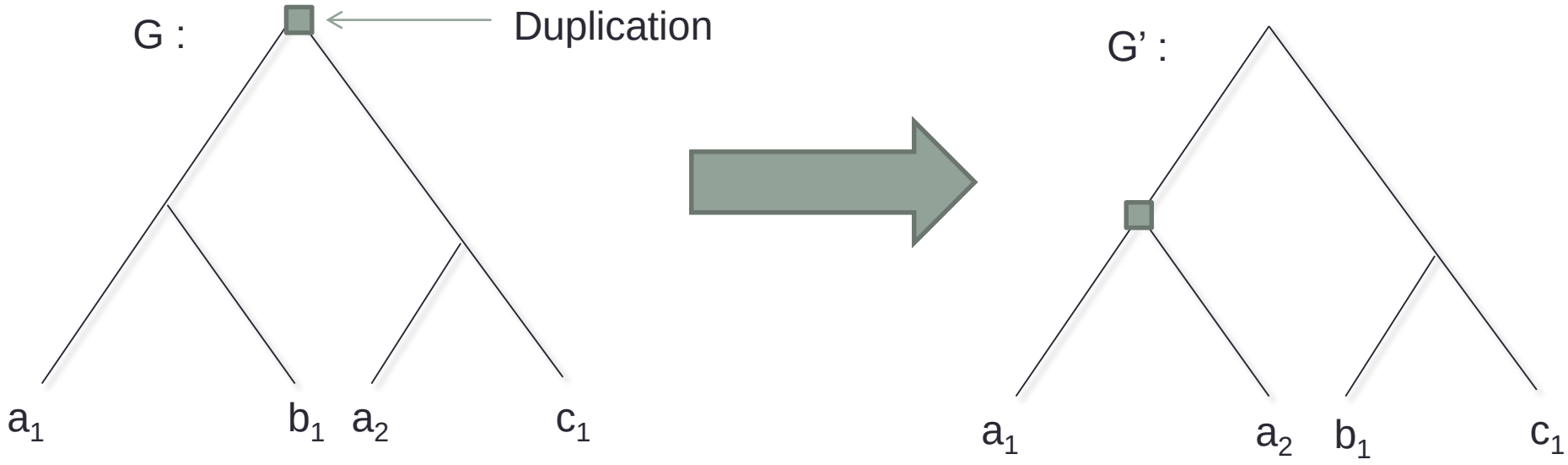
R'



Orthologs

Paralogs

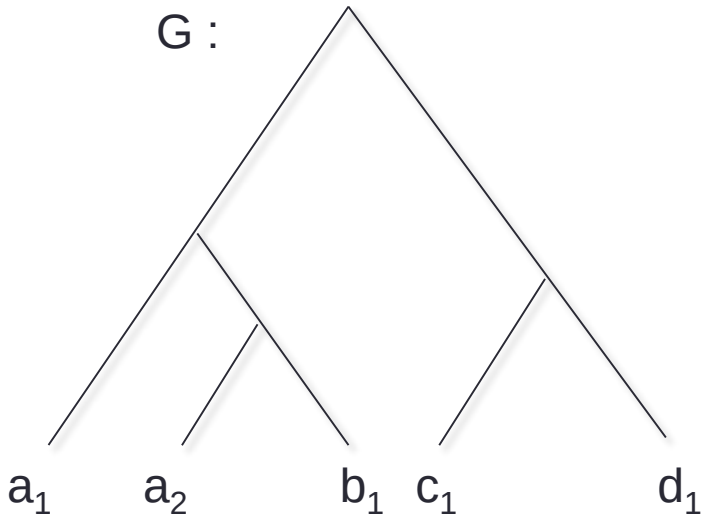
Gene tree correction



a_1 and c_1 should be orthologs.
Please do something about it…

Introduction

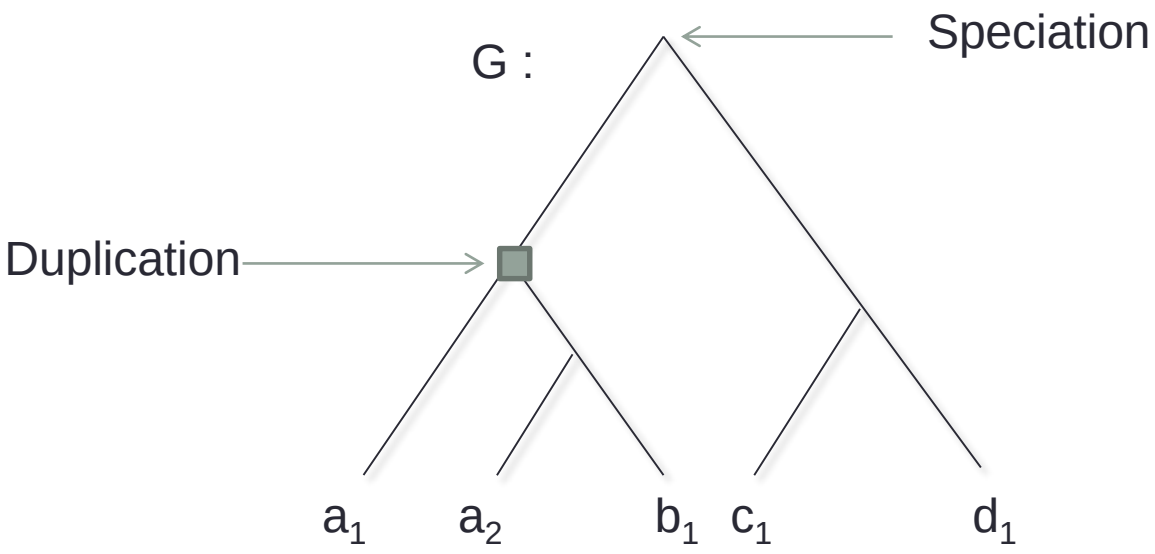
- **Gene trees** reflect the evolutionary history of a family of **homologous** genes
  - Genes that all descend from a common ancestor



a,b,c,d are species

Introduction

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



Introduction

Orthologs : LCA has undergone speciation

Paralogs : LCA has undergone duplication

(LCA = Lowest Common Ancestor)

Introduction

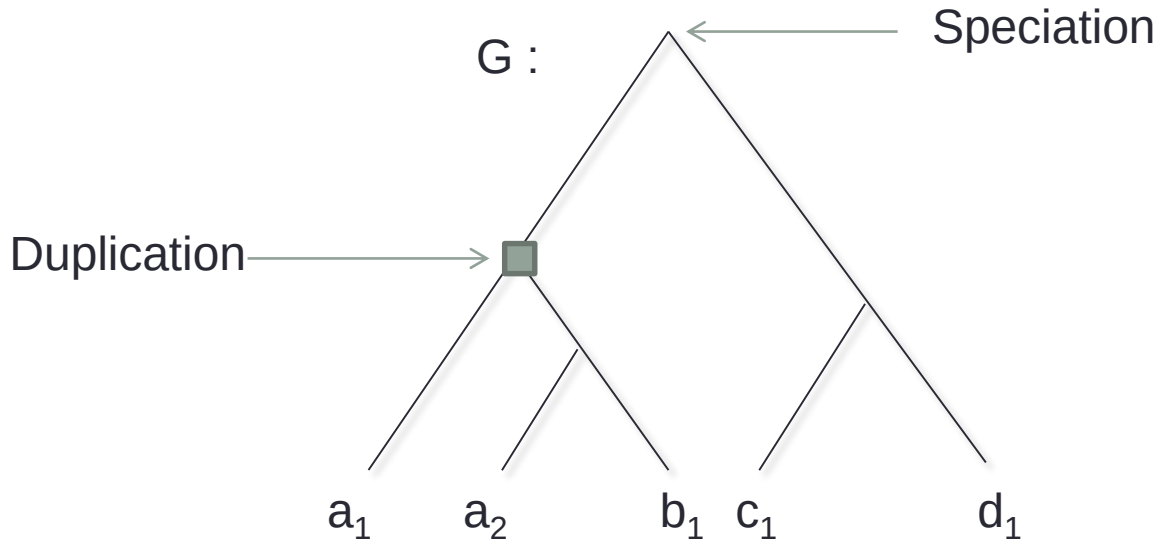
Orthologs : LCA has undergone speciation

Paralogs : LCA has undergone duplication

(LCA = Lowest Common Ancestor)

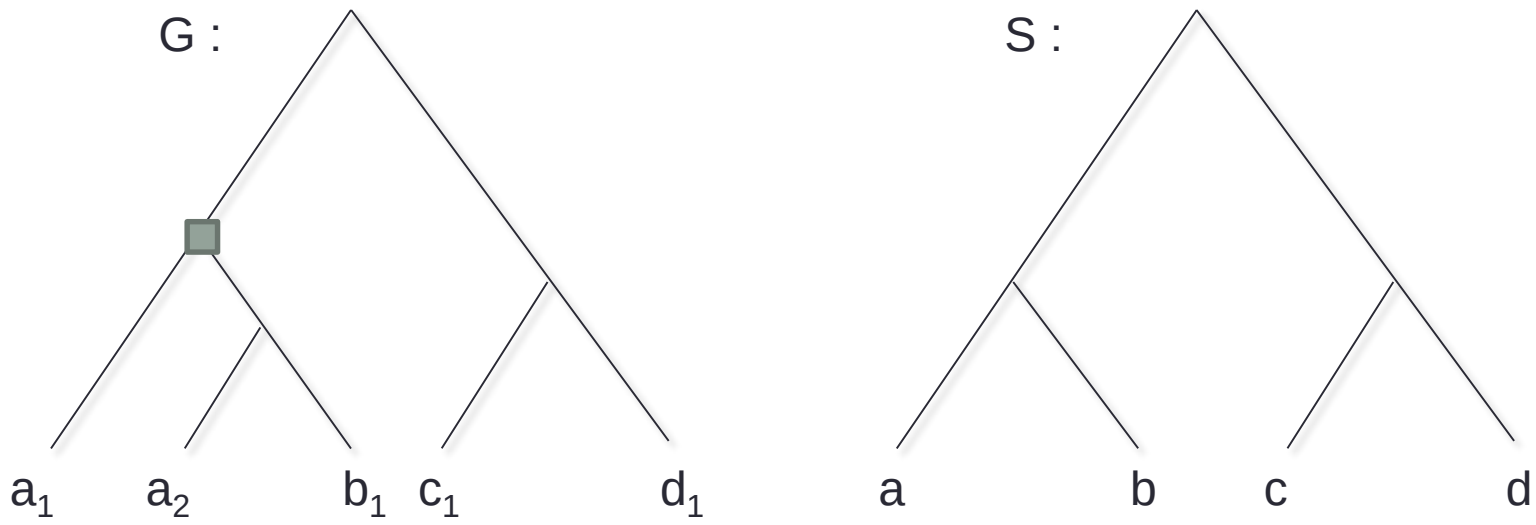
For instance, **according to G** :

- a_1, b_1 are paralogs
- a_1, c_1 are orthologs



Introduction

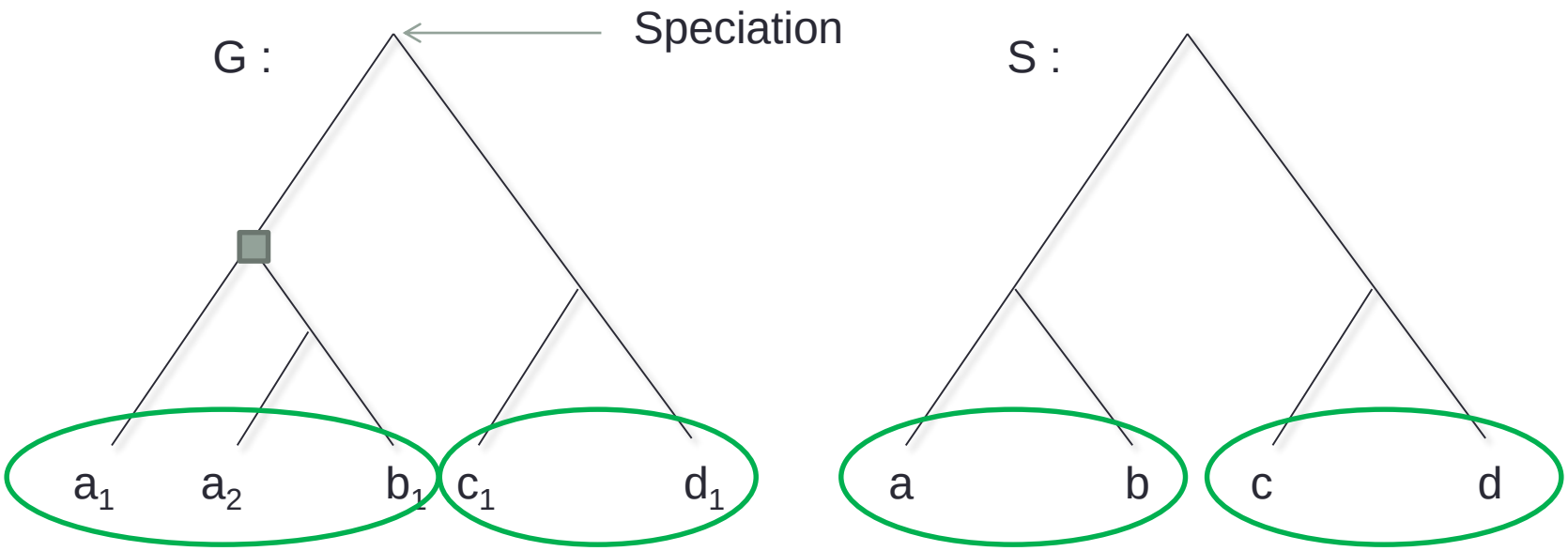
S-Consistency : If genes from species sets X,Y are separated by speciation in G, then species X, Y are separated in the **species tree** S.



Introduction

S-Consistency : If genes from species sets X,Y are separated by speciation in G, then species X, Y are separated in the **species tree** S.

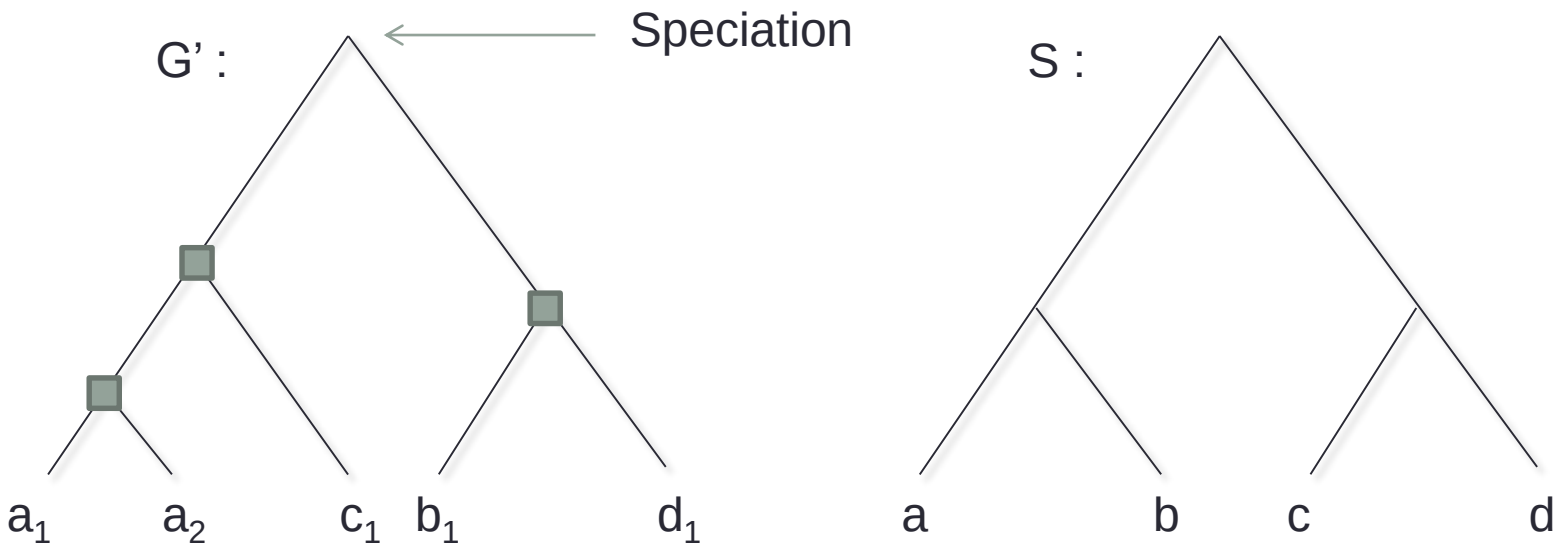
S-CONSISTENT



Introduction

S-Consistency : If genes from species sets X,Y are separated by speciation in G, then species X, Y are separated in the **species tree** S.

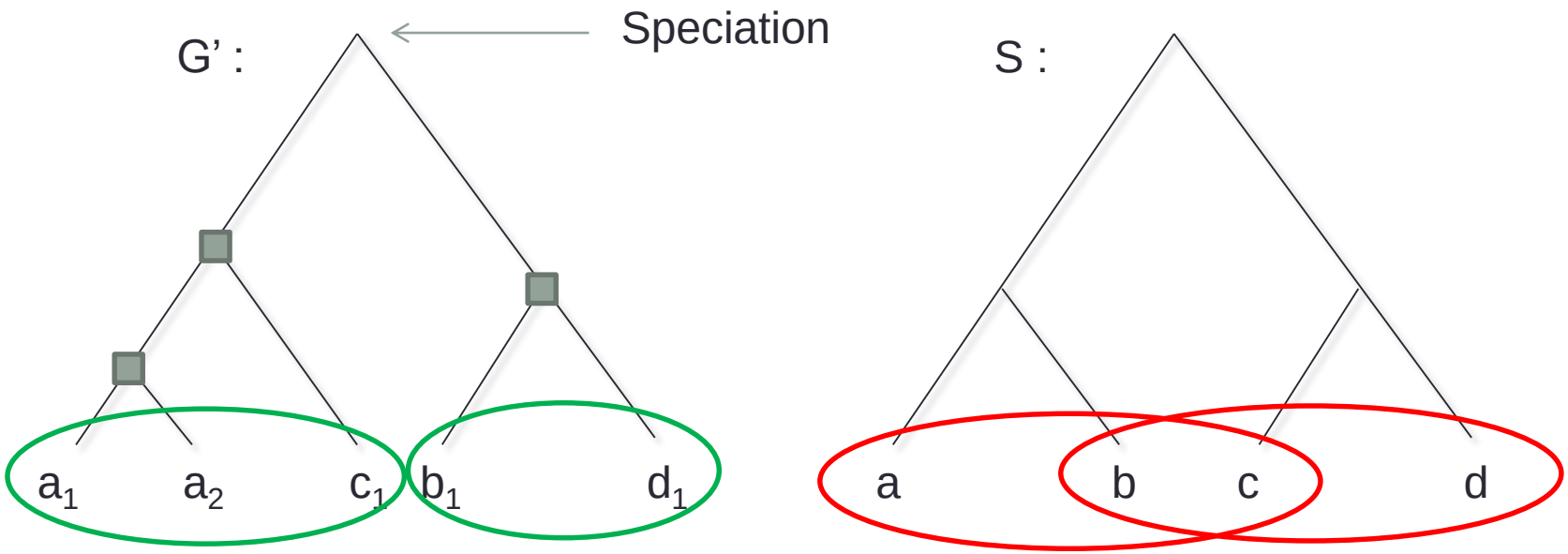
NOT S-CONSISTENT



Introduction

S-Consistency : If genes from species sets X,Y are separated by speciation in G, then species X, Y are separated in the **species tree** S.

NOT S-CONSISTENT



Magic orthology / paralogy machine

Orthologs = (a,b) (a, c) (b, c) (c, d)

Paralogs = (a, d) (b, d)

Magic orthology / paralogy machine

Orthologs = (a,b) (a, c) (b, c) (c, d)

Paralogs = (a, d) (b, d)

Sequence-based : COG, OrthoMCL, InParanoid, Proteinortho, …

Gene order/synteny-based : GIGA, SYNERGY, …

Satisfiability

Orthologs = (a,b) (a, c) (b, c) (c, d)

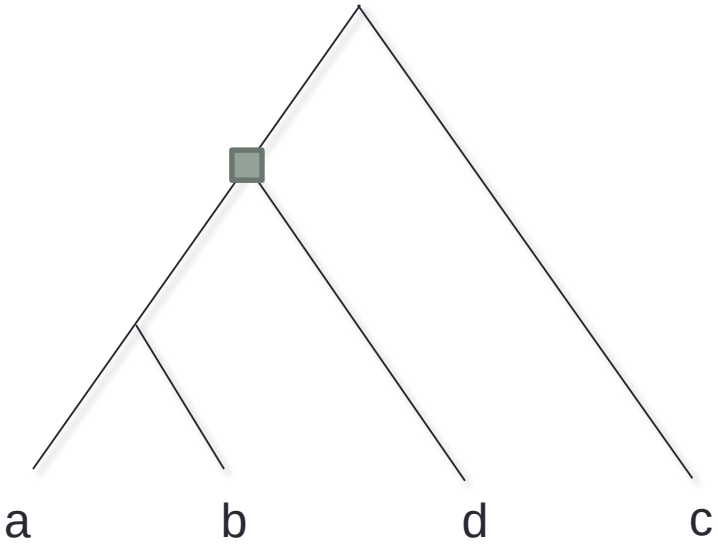
Paralogs = (a, d) (b, d)

Is there some **gene tree** and
Dup/Spec labeling that displays
these relationships ?

Satisfiability

Orthologs = (a,b) (a, c) (b, c) (c, d)

Paralogs = (a, d) (b, d)



Satisfiability

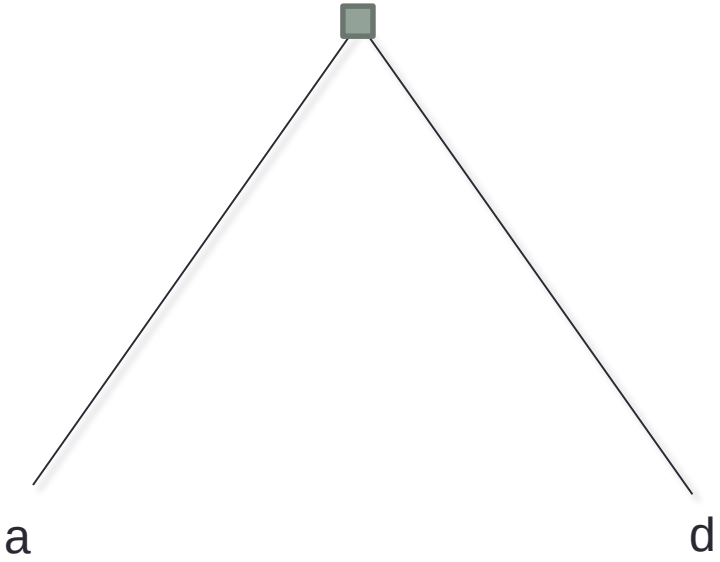
Orthologs = (a,b) (a, c) (c, d)

Paralogs = (a, d) (b, c) (b, d)

Satisfiability

Orthologs = (a,b) (a, c) (c, d)

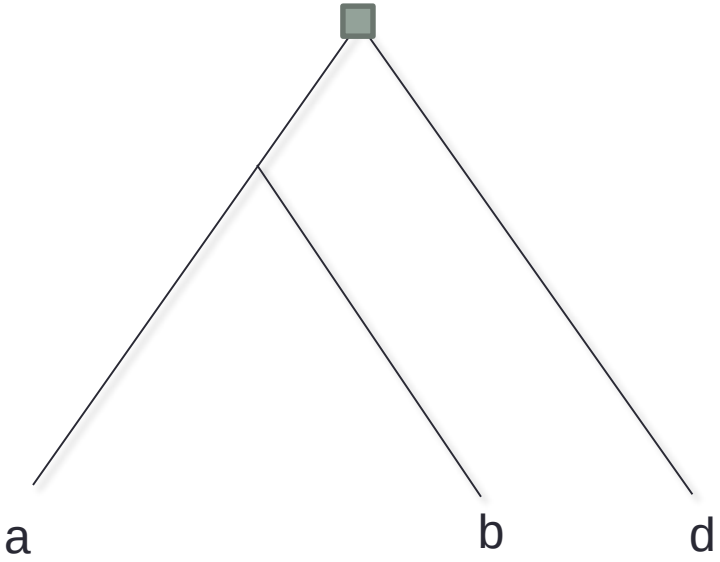
Paralogs = (a, d) (b, c) (b, d)



Satisfiability

Orthologs = (a,b) (a, c) (c, d)

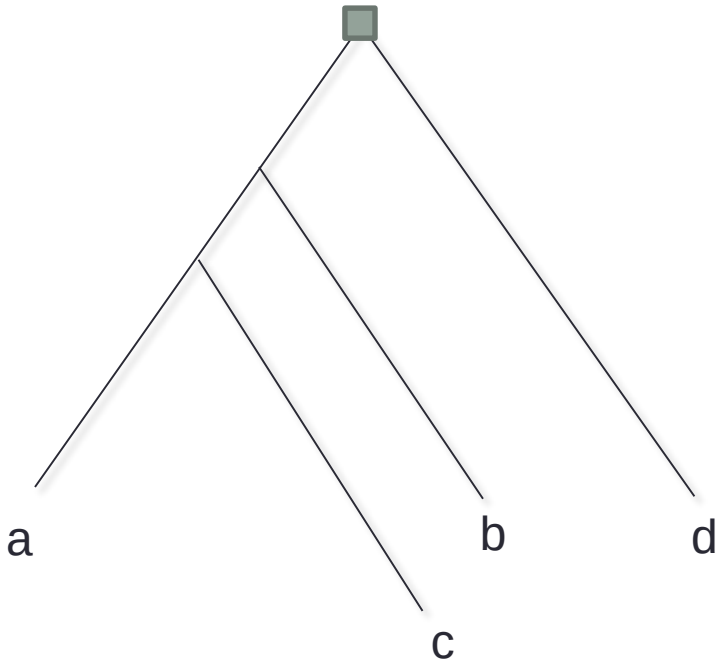
Paralogs = (a, d) (b, c) (b, d)



Satisfiability

Orthologs = (a,b) (a, c) (c, d)

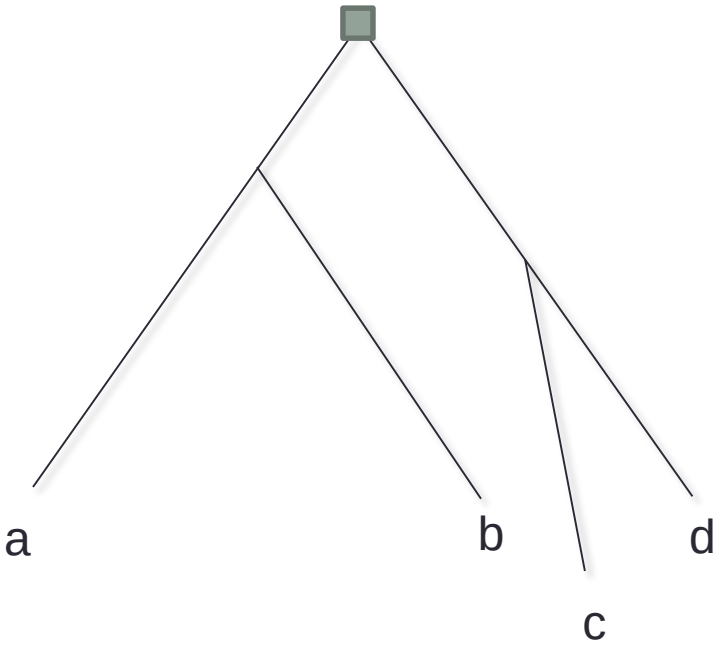
Paralogs = (a, d) (b, c) (b, d)



Satisfiability

Orthologs = (a,b) (a, c) (c, d)

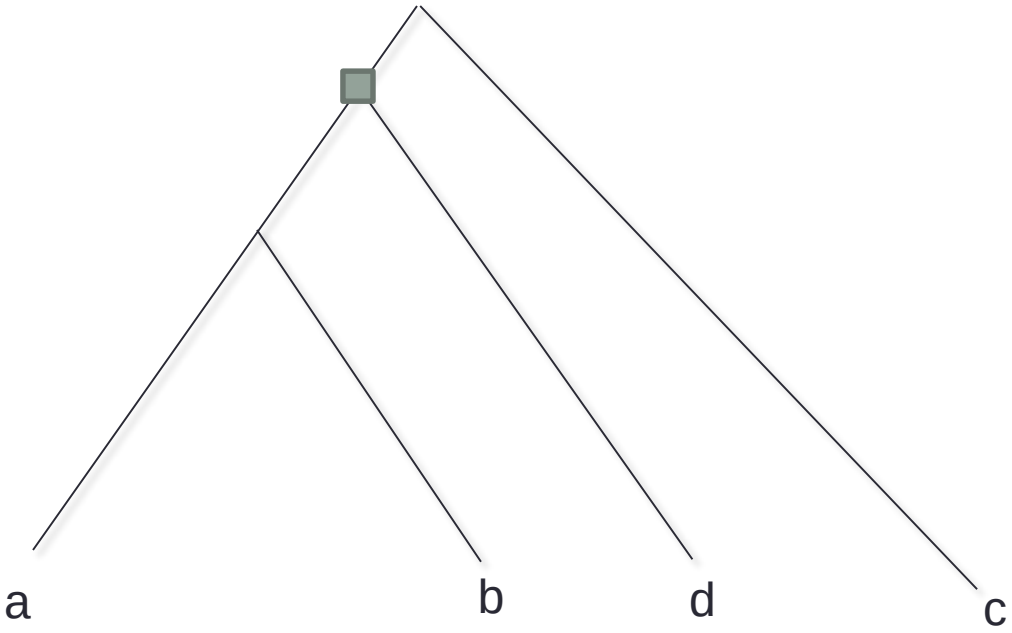
Paralogs = (a, d) (b, c) (b, d)



Satisfiability

Orthologs = (a,b) (a, c) (c, d)

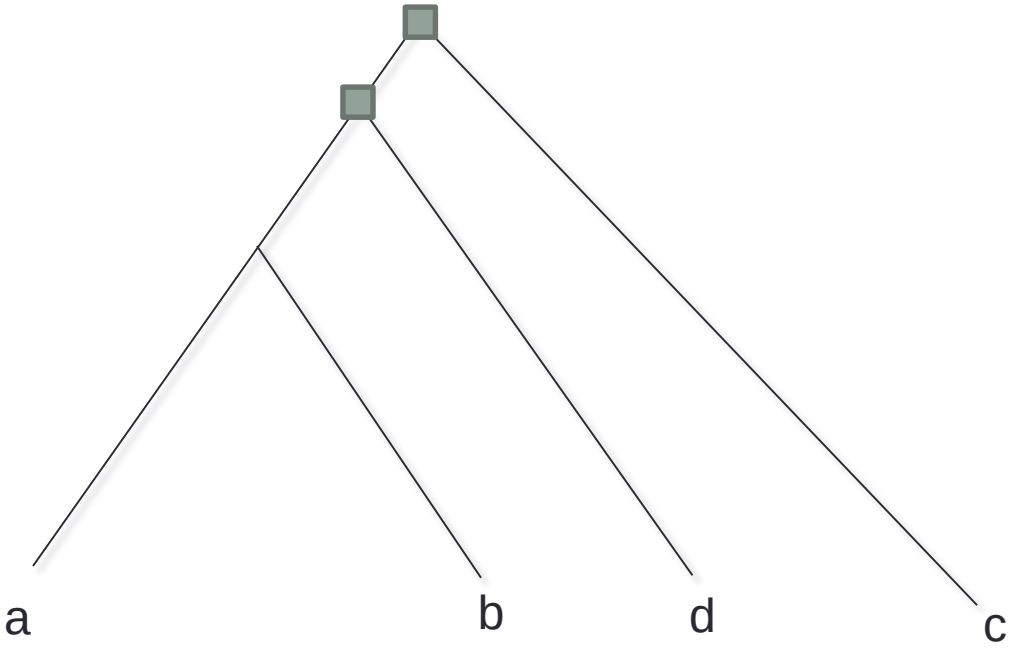
Paralogs = (a, d) (b, c) (b, d)



Satisfiability

Orthologs = (a,b) (a, c) (c, d)

Paralogs = (a, d) (b, c) (b, d)



Satisfiability

Orthologs = (a,b) (a, c) (c, d)

Paralogs = (a, d) (b, c) (b, d)

I JUST CAN'T !

THESE DON'T MAKE SENSE !

S-Consistency

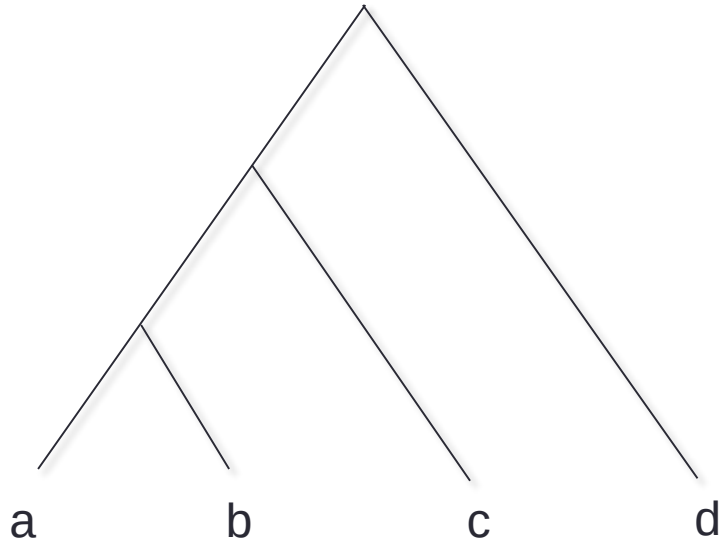
Orthologs = (a,d) (c, d)

Paralogs = (a, b) (a, c) (b, c) (b, d)

Gene tree G

?

Species tree S

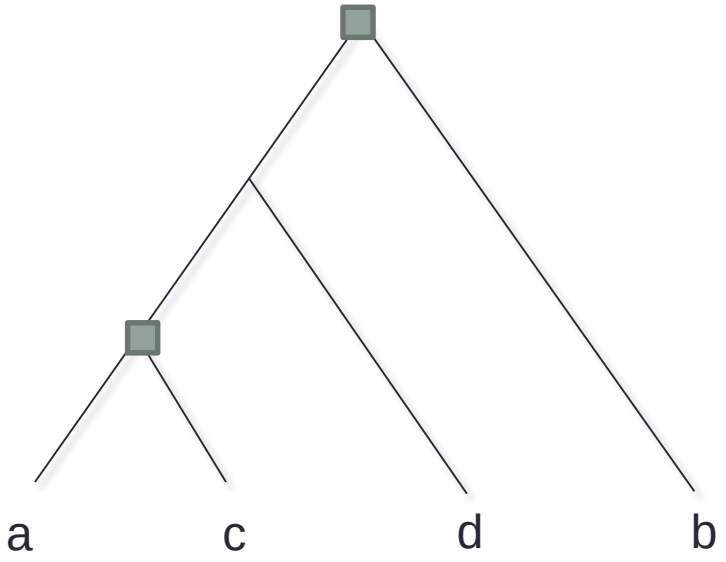


S-Consistency

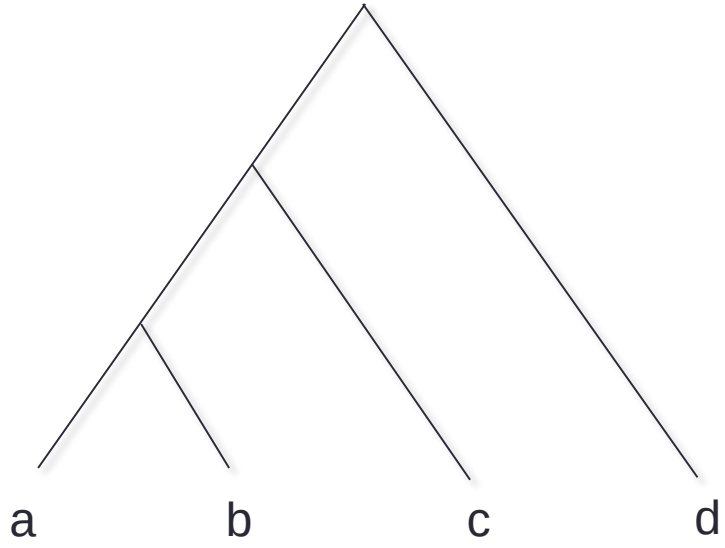
Orthologs = (a,d) (c, d)

Paralogs = (a, b) (a, c) (b, c) (b, d)

Gene tree G



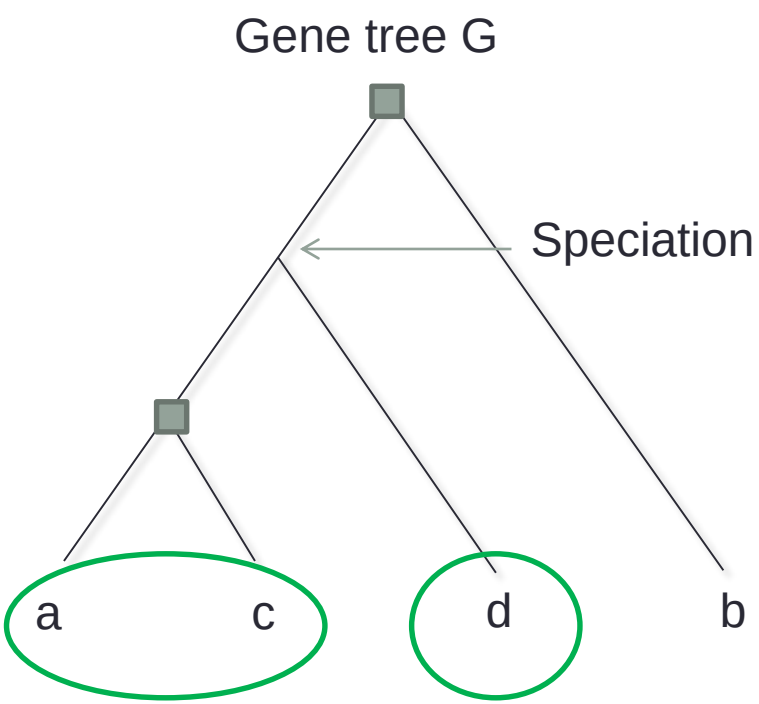
Species tree S

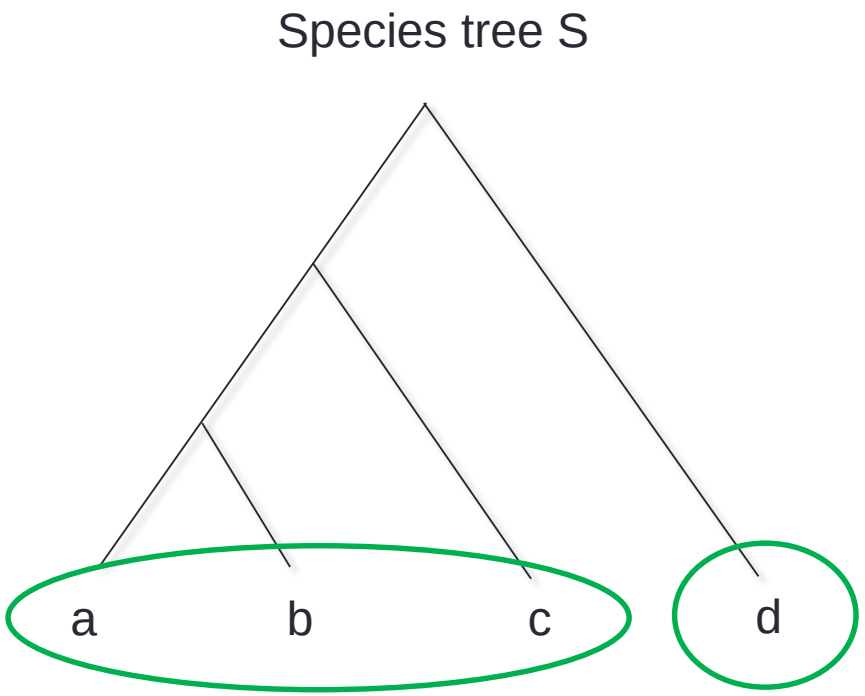


S-Consistency

Orthologs = (a,d) (c, d)

Paralogs = (a, b) (a, c) (b, c) (b, d)





S-Consistency

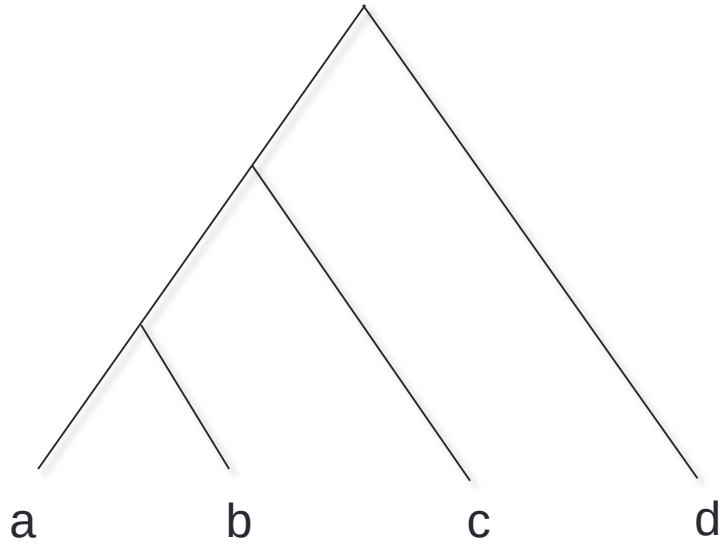
Orthologs = (a,b) (a, d) (b, c) (c, d)

Paralogs = (a, c) (b, d)

Gene tree G

?

Species tree S

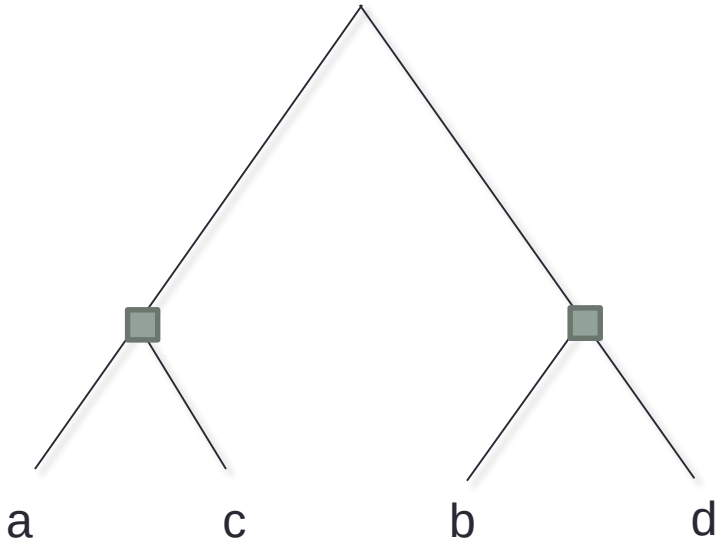


S-Consistency

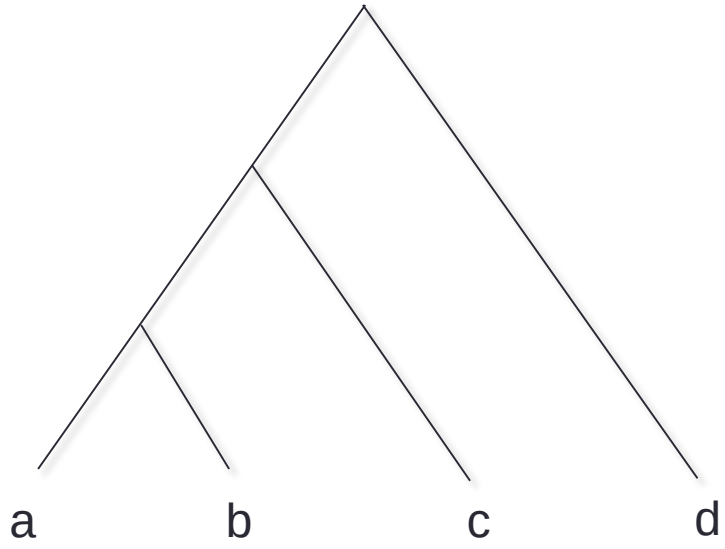
Orthologs = (a,b) (a, d) (b, c) (c, d)

Paralogs = (a, c) (b, d)

Gene tree G



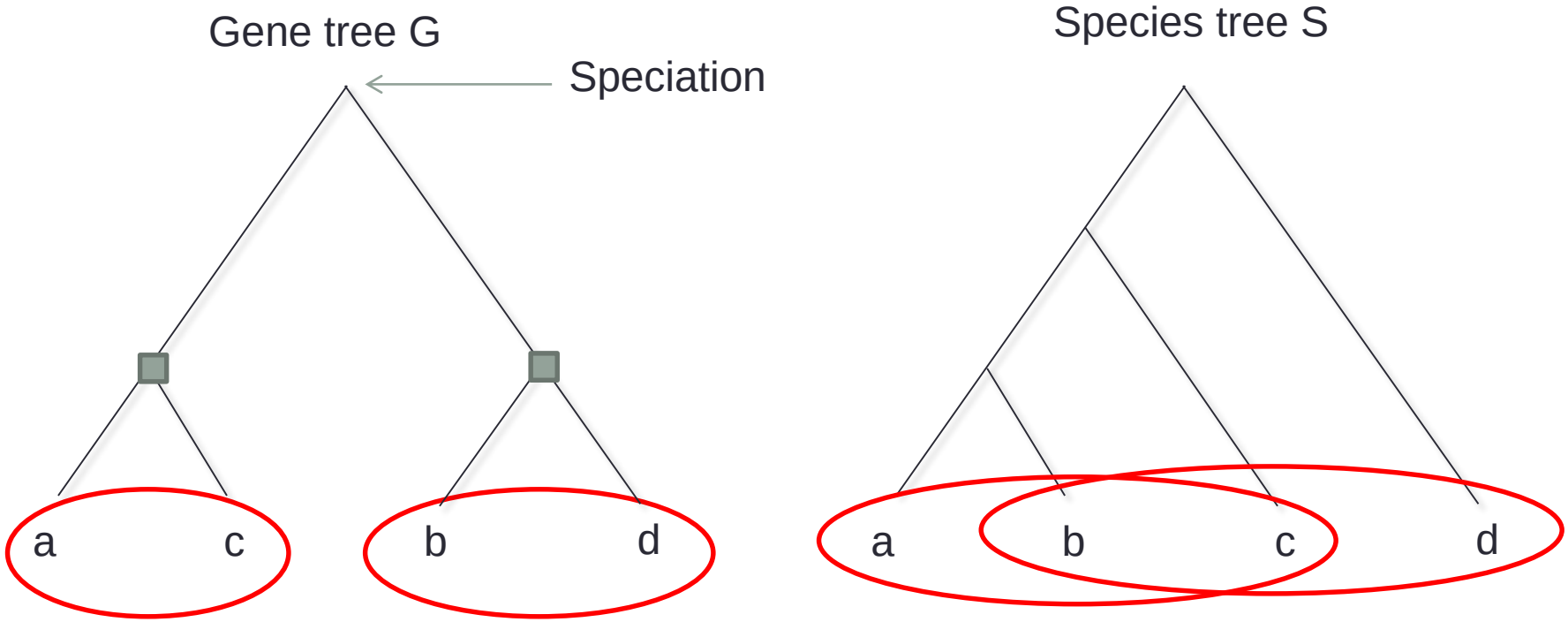
Species tree S



S-Consistency

Orthologs = (a,b) (a, d) (b, c) (c, d)

Paralogs = (a, c) (b, d)



The point is

Some people / software infer **orthology/paralogy** relations.

The point is

Some people / software infer **orthology/paralogy** relations.

Do they make any sense ?

The point is

Some people / software infer **orthology/paralogy** relations.

Do they make any sense ?

- Is there **some gene tree** that displays them ?

The point is

Some people / software infer **orthology/paralogy** relations.

Do they make any sense ?

- Is there **some gene tree** that displays them ? (not always)

The point is

Some people / software infer **orthology/paralogy** relations.

Do they make any sense ?

- Is there **some gene tree** that displays them ? (not always)
- Out of these, is there one that **agrees with our species tree** ?

The point is

Some people / software infer **orthology/paralogy** relations.

Do they make any sense ?

- Is there **some gene tree** that displays them ? (not always)
- Out of these, is there one that **agrees with our species tree** ? (not always)

The point is

Some people / software infer **orthology/paralogy** relations.

Do they make any sense ?

- Is there **some gene tree** that displays them ? (not always)
- Out of these, is there one that **agrees with our species tree** ? (not always)

If they don’t make sense, what should we do about it ?

Problem 1

Given : a species tree S and a set of orthology / paralogy relations R.

Find : the minimum number of relations to change in R such that the relations are satisfiable + S-Consistent

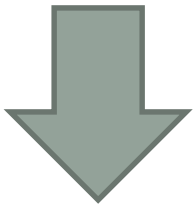
Problem 1

Given : a species tree S and a set of orthology / paralogy relations R.

Find : the minimum number of relations to change in R such that the relations are satisfiable + S-Consistent

Orthologs = (a,b) (a, d) (b, c) (c, d)

Paralogs = (a, c) (b, d)



Orthologs = ~~***(a,b)***~~ (a, d) (b, c) ~~***(c, d)***~~

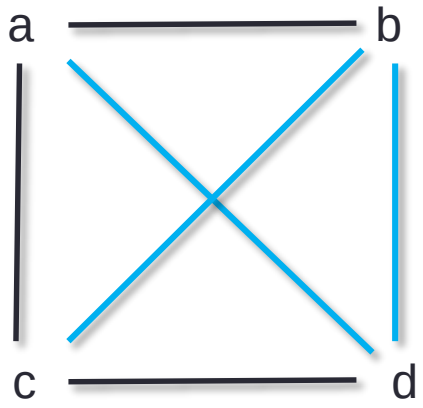
Paralogs = (a, c) (b, d) **(a,b) (c, d)**

Relation graphs

Orthologs = (a,b) (a, c) (c, d)

Paralogs = (a, d) (b, c) (b, d)

Relation graph R



Orthologs

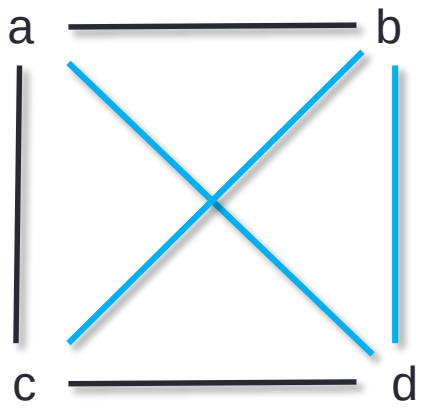
Paralogs

Relation graphs

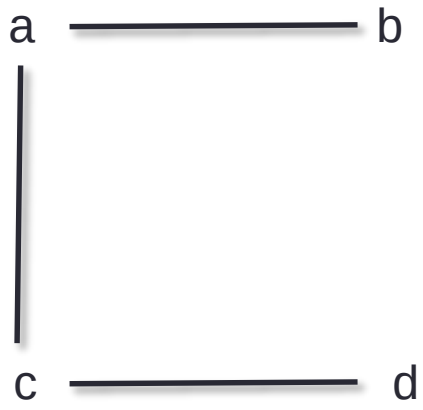
Orthologs = (a,b) (a, c) (c, d)

Paralogs = (a, d) (b, c) (b, d)

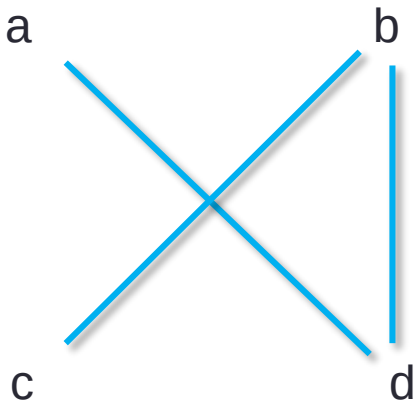
R



R_O



R_P



Orthologs

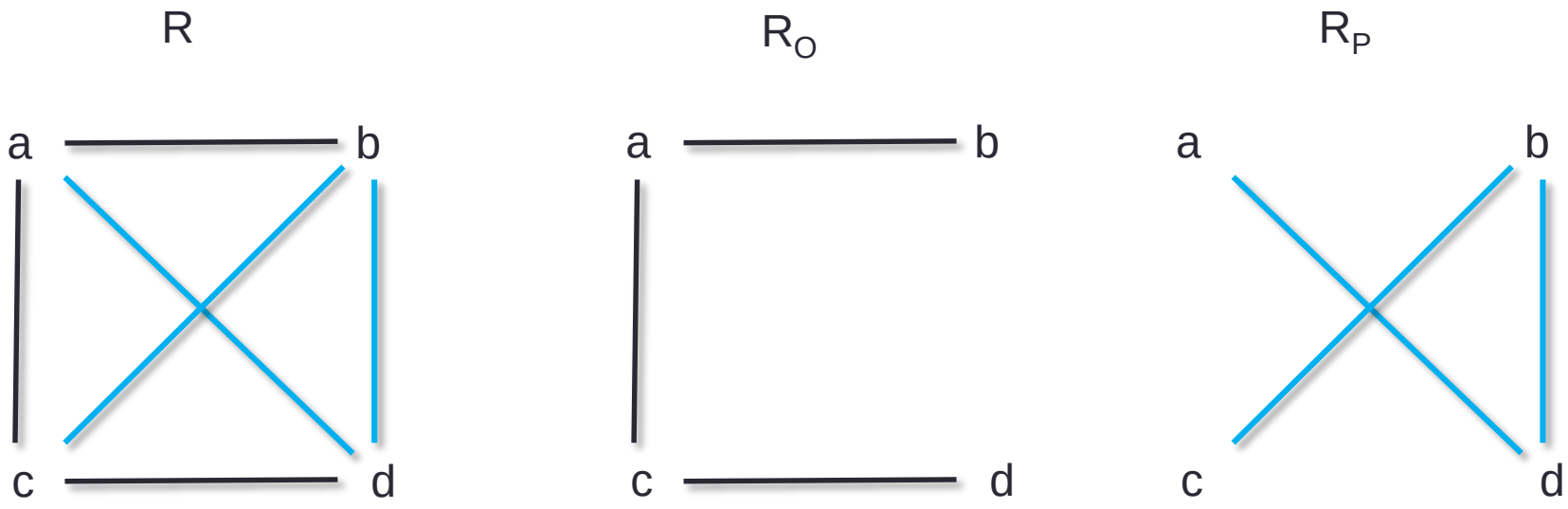
Paralogs

Relation graphs

(Hernandez-Rosales & al., 2012)

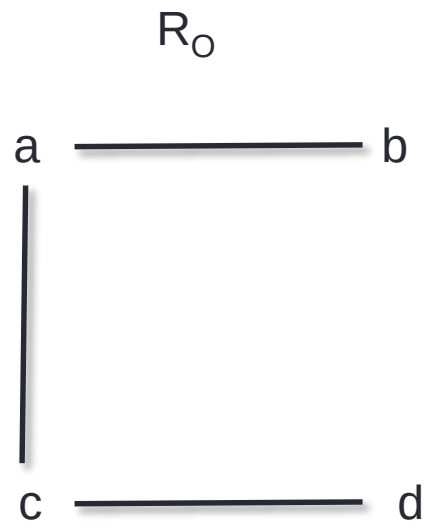
The relations are **satisfiable** iff

R_O is P_4-free (no **induced path** of length 4)



Relation graphs

NOT SATISFIABLE : No gene tree displays these relations



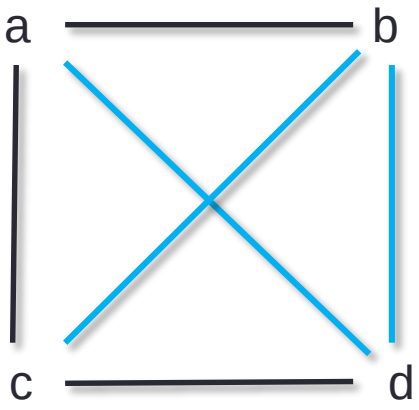
Orthologs

Paralogs

Relation graphs

Makes no sense

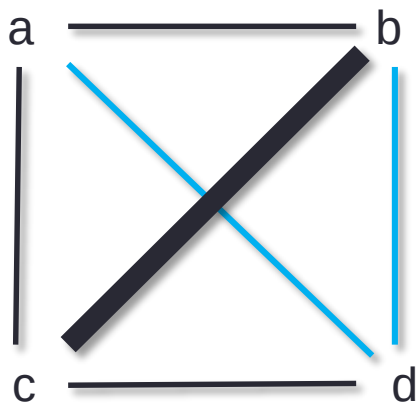
R



Correction

Makes sense

R’

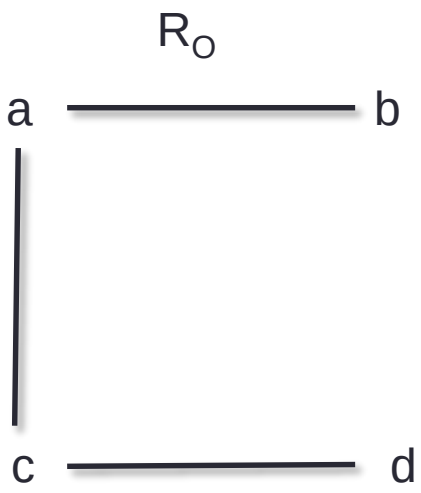


Orthologs

Paralogs

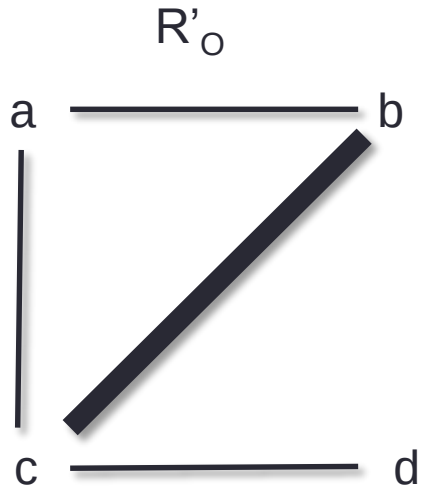
Relation graphs

Makes no sense



Correction

Makes sense



Orthologs

Paralogs

Problem 1

Given : a species tree S and a set of orthology / paralogy relations R.

Find : the minimum number of relations to change in R such that the relations are **satisfiable** ~~+ S-Consistent~~

Problem 1

Given : a species tree S and a set of orthology / paralogy relations R.

Find : the minimum number of relations to change in R such that the relations are **satisfiable** ~~+ S-Consistent~~

Equivalently : find the minimum number of edge insertions / deletions in R_O such that it is P_4-free.

Problem 1

Given : a species tree S and a set of orthology / paralogy relations R.

Find : the minimum number of relations to change in R such that the relations are **satisfiable** ~~+ S-Consistent~~

Equivalently : find the minimum number of edge insertions / deletions in R_O such that it is P_4-free.

NP-COMPLETE (El-Mallah and Colbourn, 1988)

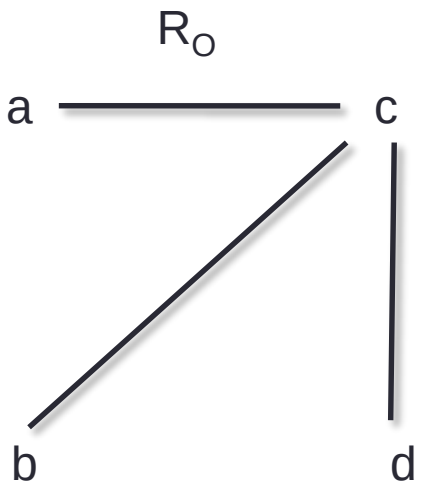
Problem 1

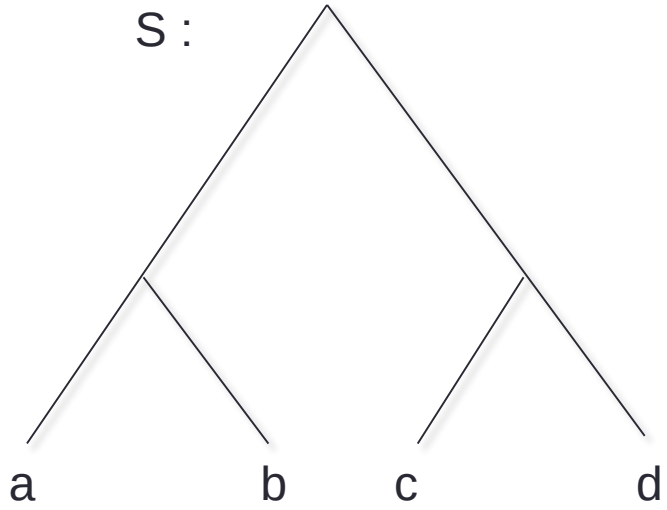
Given : a species tree S and a set of orthology / paralogy relations R.

Find : the minimum number of relations to change in R such that the relations are **satisfiable** + **S-Consistent**

What about S-Consistency ???

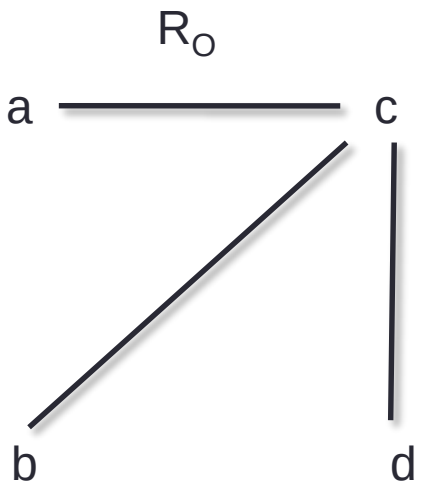
Relation graphs and S-Consistency

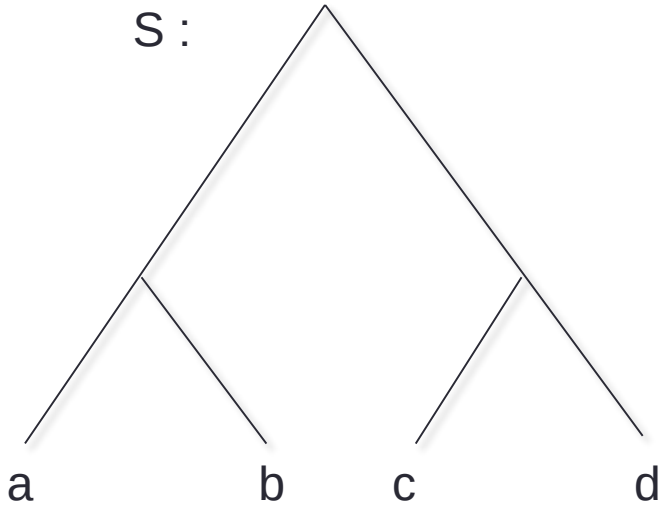




Does R_O lead to an S-Consistent gene tree ?

Relation graphs and S-Consistency



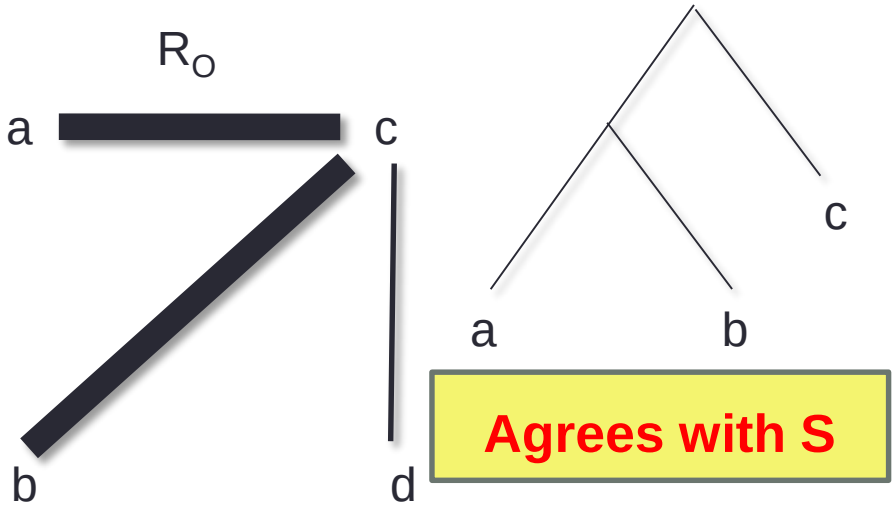


Does R_O lead to an S-Consistent gene tree ?

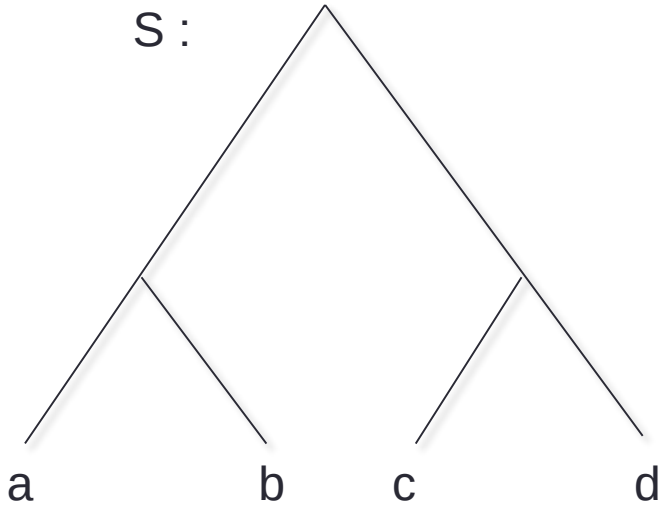
Each **P_3** represents a mandatory **speciation triplet**

(i.e. a speciation separating two species from another).

Relation graphs and S-Consistency



Agrees with S

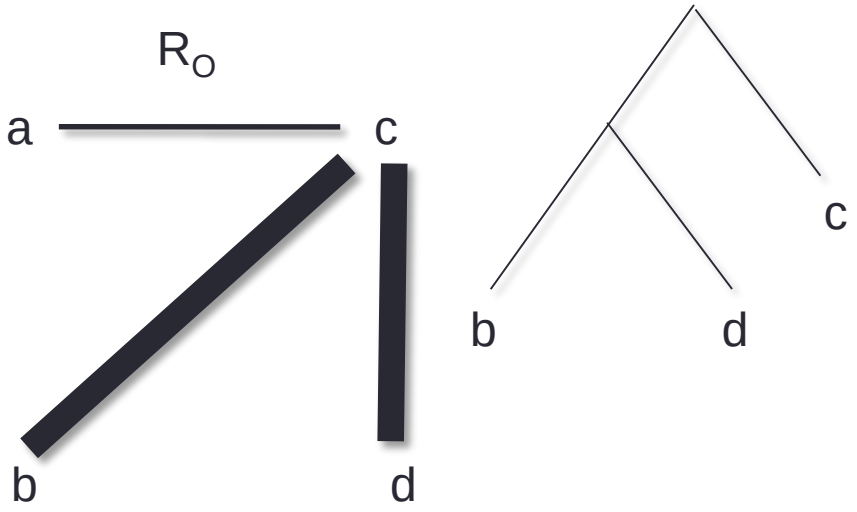


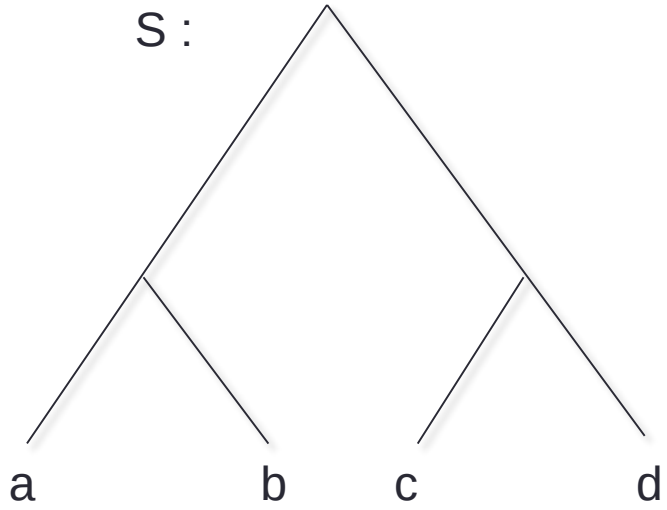
Does R_O lead to an S-Consistent gene tree ?

Each **P_3** represents a mandatory **speciation triplet**

(i.e. a speciation separating two species from another).

Relation graphs and S-Consistency



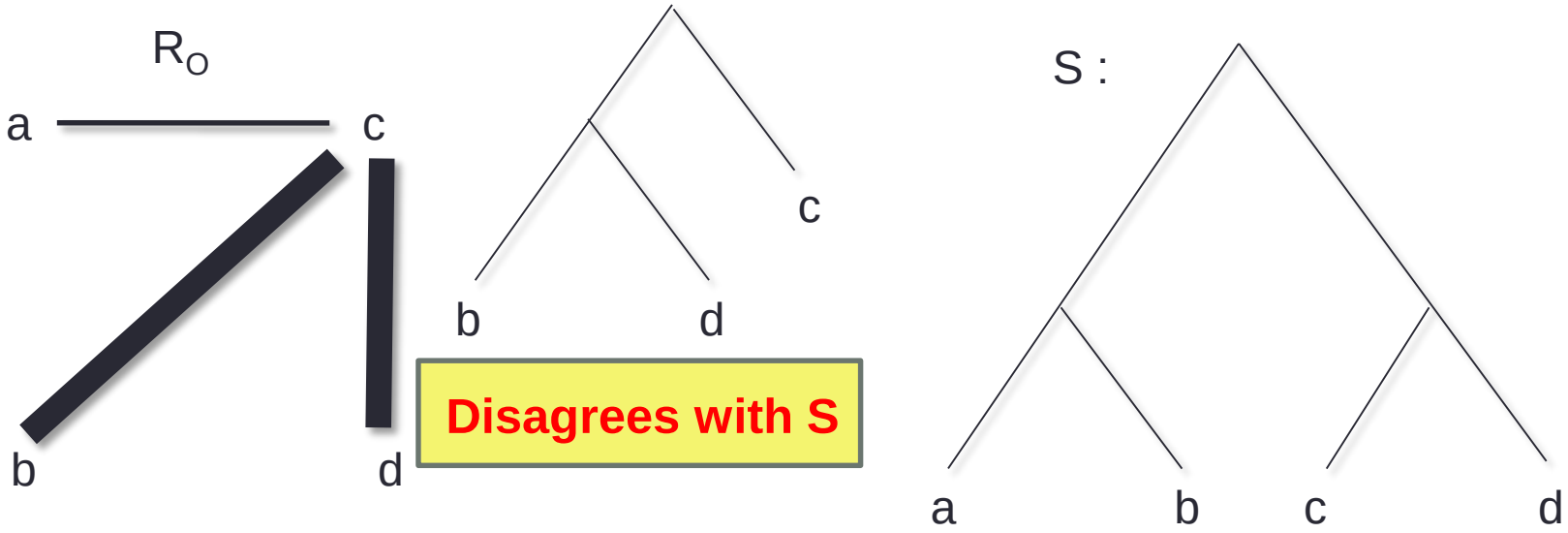


Does R_O lead to an S-Consistent gene tree ?

Each **P_3** represents a mandatory **speciation triplet**

(i.e. a speciation separating two species from another).

Relation graphs and S-Consistency



Does R_O lead to an S-Consistent gene tree ?

Each **P_3** represents a mandatory **speciation triplet**

(i.e. a speciation separating two species from another).

Relation graphs and S-Consistency

The (satisfiable) relations are **S-Consistent** iff each P_3 of R_O represents a triplet that agrees with S.

Relation graphs and S-Consistency

The (satisfiable) relations are **S-Consistent** iff

each P_3 of R_O represents a triplet that agrees with S.

Checking for S-Consistency is easy. Just find every P_3 and check them, in $O(n^3)$.

But how to correct for S-Consistency ?

Problem 1

Given : a species tree S and a set of orthology / paralogy relations R.

Find : the minimum number of relations to change in R such that the relations are **satisfiable** + **S-Consistent**

Equivalently : find the minimum number of edge insertions / deletions in R_O such that it is P_4-free + every P_3 agrees with S.

Problem 1

Given : a species tree S and a set of orthology / paralogy relations R.

Find : the minimum number of relations to change in R such that the relations are **satisfiable** + **S-Consistent**

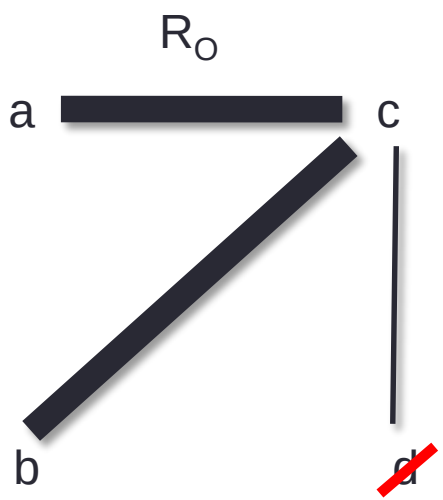
Equivalently : find the minimum number of edge insertions / deletions in R_O such that it is P_4-free + every P_3 agrees with S.

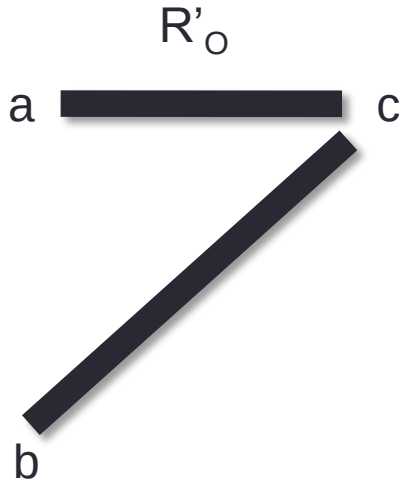
NP-COMPLETE (Lafond and El-Mabrouk, 2015)

Problem 2

Given : a species tree S and a set of orthology / paralogy relations R.

Find : the minimum number of **genes to remove** such that the relations are **satisfiable** + **S-Consistent**





Problem 2

Given : a species tree S and a set of orthology / paralogy relations R.

Find : the minimum number of **genes to remove** such that the relations are **satisfiable** + **S-Consistent**

NP-COMPLETE (Lafond and El-Mabrouk, 2015)

Problem 2

Given : a species tree S and a set of orthology / paralogy relations R.

Find : the minimum number of **genes to remove** such that the relations are **satisfiable** + **S-Consistent**

NP-COMPLETE (Lafond and El-Mabrouk, 2015)

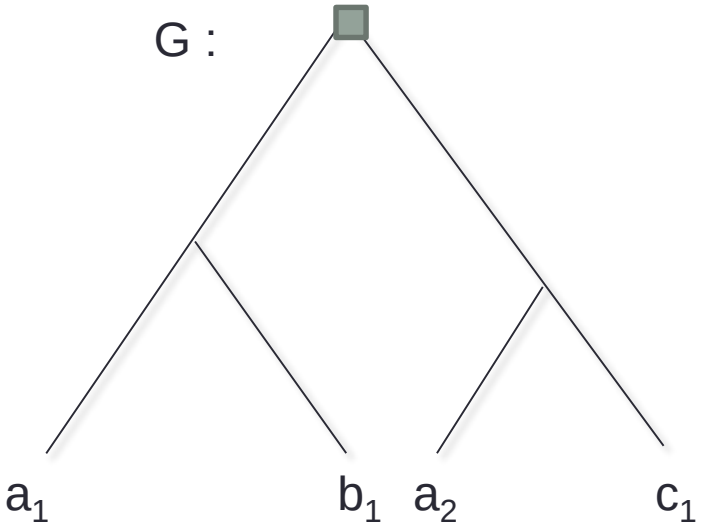
WORSE ! Hard to approximate within a $n^{1-\varepsilon}$ factor (R.Dondi, unpublished).

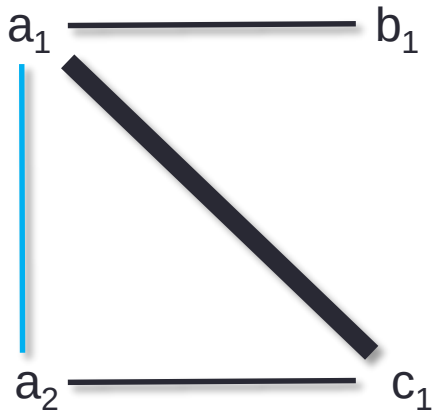
Allright then…

So we can’t correct inferred relations.

Still, let’s say there’s a bunch of them that I trust.

Can I use them to validate / correct my gene tree ?

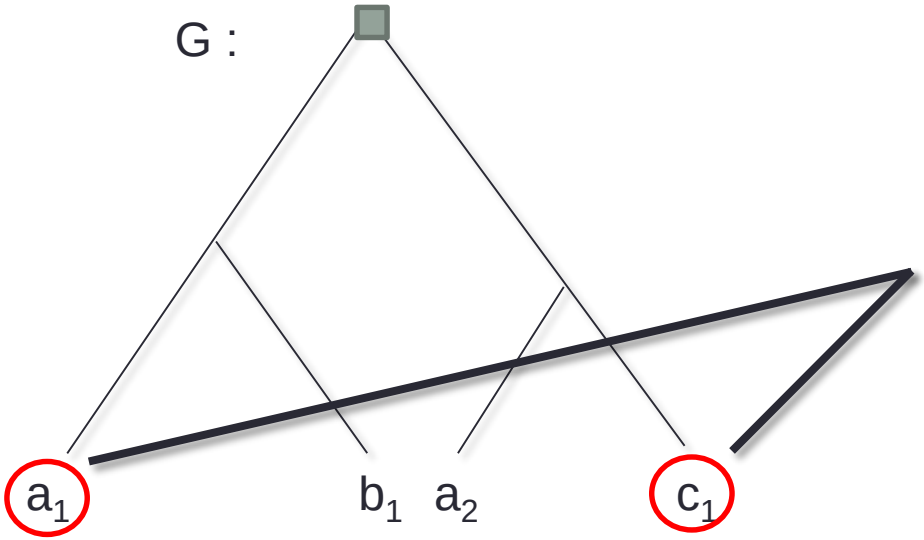




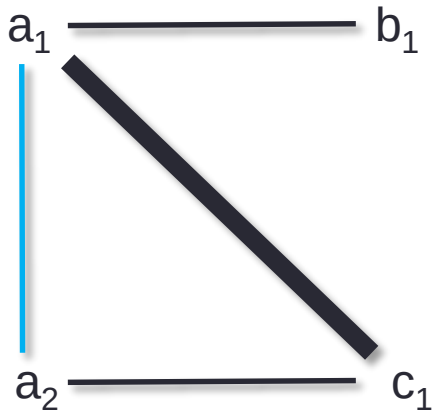
Partial, but trustworthy, relations

Orthologs

Paralogs



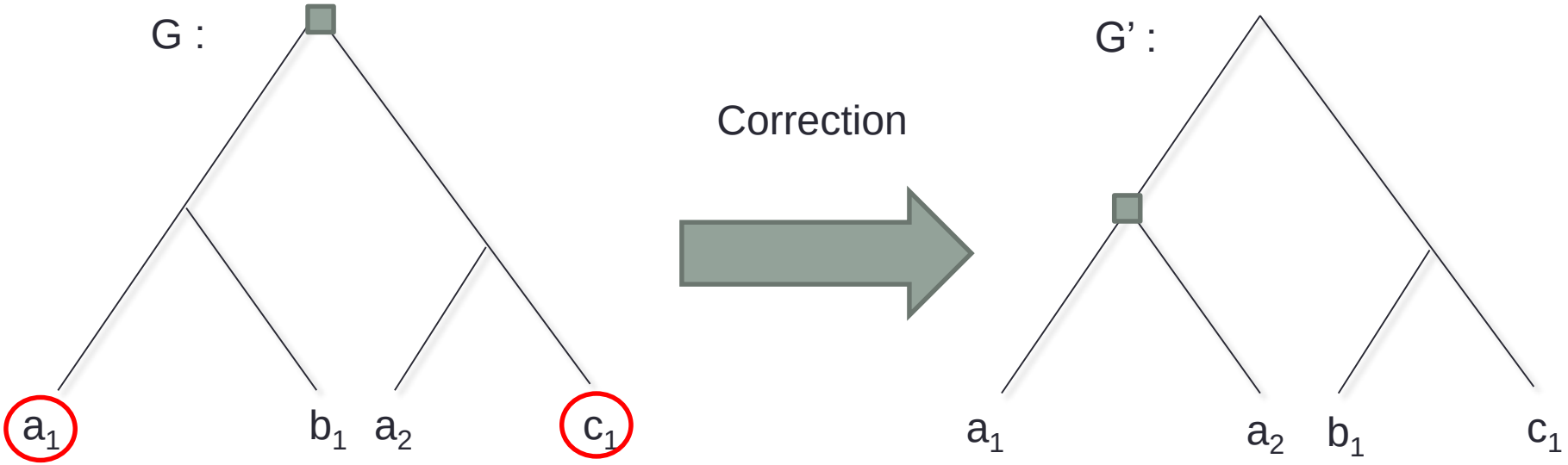
These guys should be orthologs

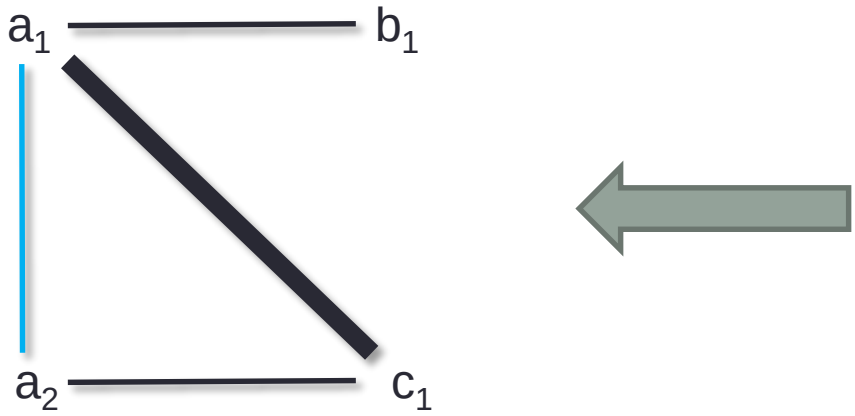


Partial, but trustworthy, relations

Orthologs

Paralogs

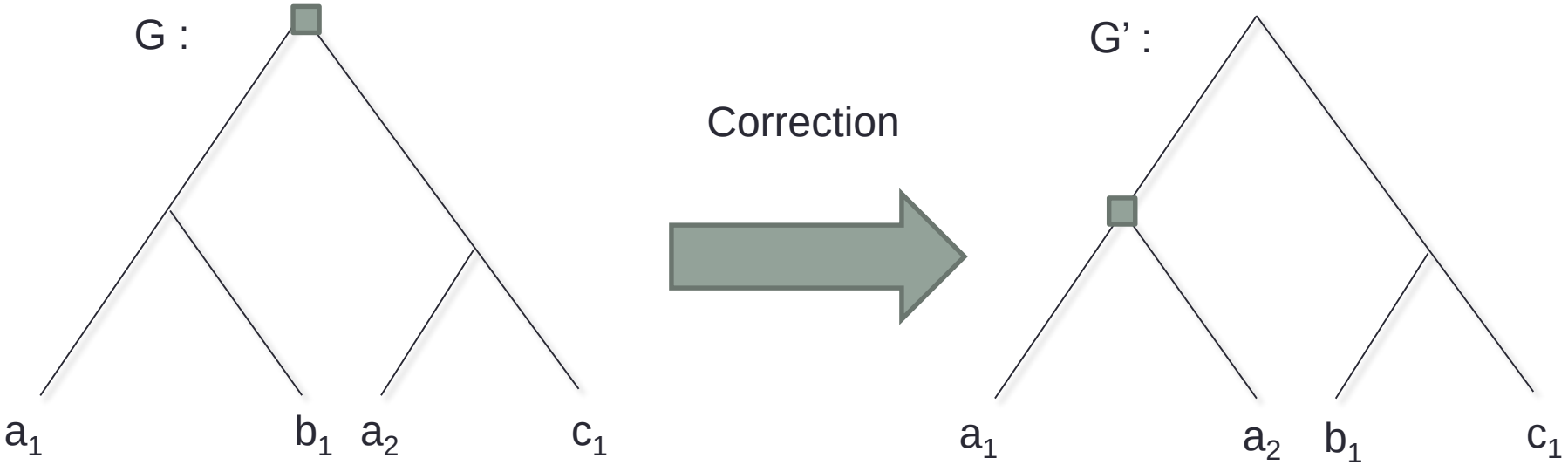




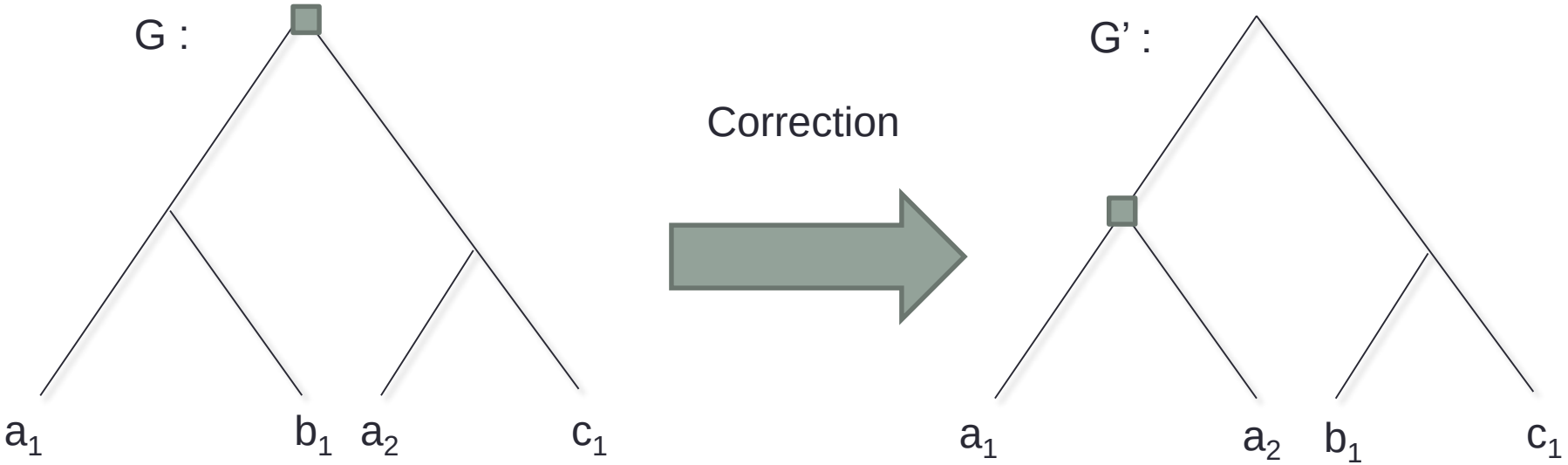
Partial, but trustworthy, relations

Orthologs

Paralogs

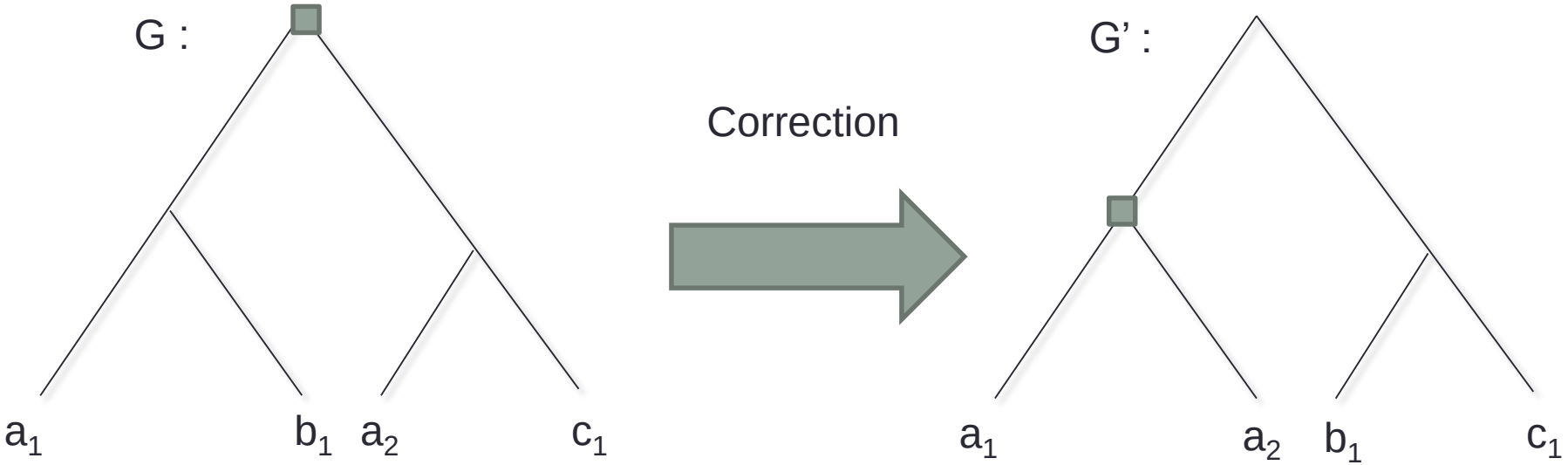


Keep G as ‘intact’ as possible.



Keep G as ‘intact’ as possible.

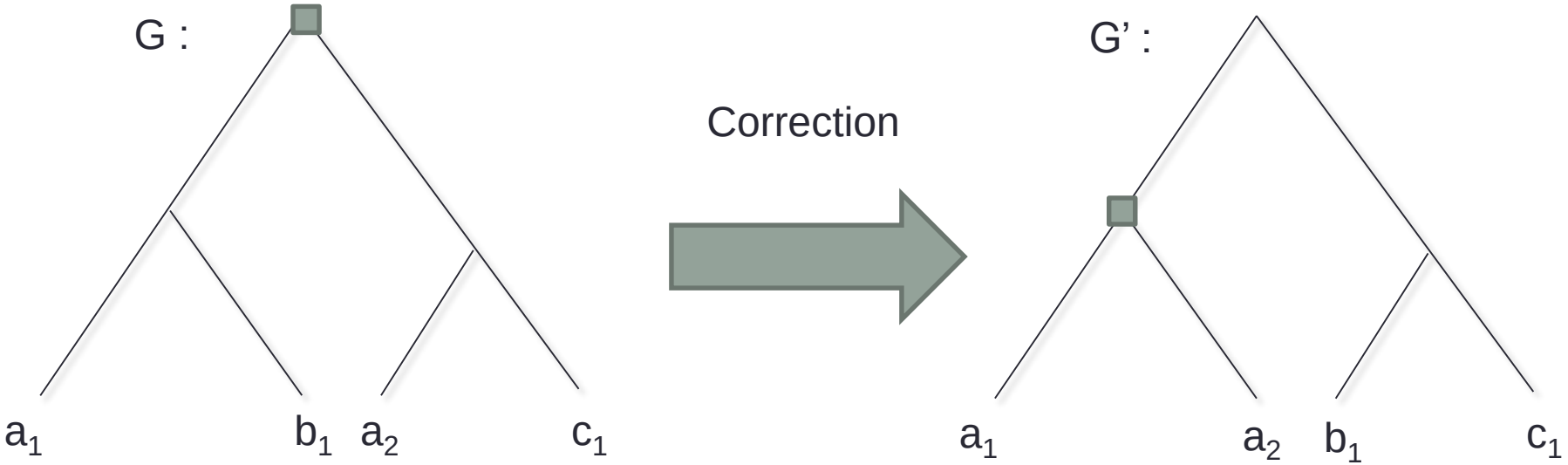
Intact 1 : minimize the number of broken relations w.r.t G.



Keep G as ‘intact’ as possible.

Intact 1 : minimize the number of broken relations w.r.t G.

For instance, here we broke the (b_1, c_1) paralogy.

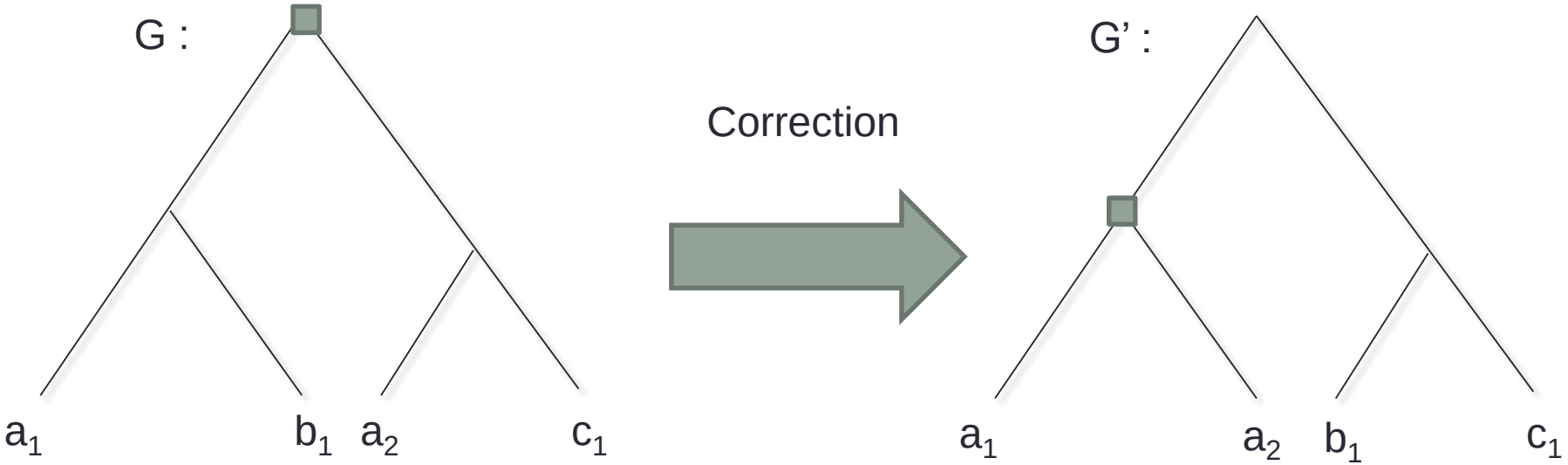


Keep G as ‘intact’ as possible.

Intact 1 : minimize the number of broken relations w.r.t G.

For instance, here we broke the (b_1, c_1) paralogy.

Intact 2 : minimize the Robinson-Foulds distance between G and G’.



Keep G as 'intact' as possible.

Intact 1 : minimize the number of broken relations w.r.t G.

For instance, here we broke the (b_1, c_1) paralogy.

Intact 2 : minimize the Robinson-Foulds distance between G and G'.

The number of clades that G and G' do not have in common.

Problem 3

Given : a species tree S, a gene tree G, and a partial set R of orthologs and paralogs to satisfy.

Find : an S-Consistent gene tree G’ that satisfies the relations of R such that…

Problem 3

Given : a species tree S, a gene tree G, and a partial set R of orthologs and paralogs to satisfy.

Find : an S-Consistent gene tree G’ that satisfies the relations of R such that…

…the number of orthologs/paralogs in G but not in G’ is minimized.

Problem 3

Given : a species tree S, a gene tree G, and a partial set R of orthologs and paralogs to satisfy.

Find : an S-Consistent gene tree G' that satisfies the relations of R such that…

…the number of orthologs/paralogs in G but not in G' is minimized.

NP-COMPLETE (Lafond and El-Mabrouk, 2015)

Reduction from the edge insertion/deletion problem (Problem 1)

Problem 4

Given : a species tree S, a gene tree G, and a partial set R of orthologs and paralogs to satisfy.

Find : an S-Consistent gene tree G’ that satisfies the relations of R such that…

Problem 4

Given : a species tree S, a gene tree G, and a partial set R of orthologs and paralogs to satisfy.

Find : an S-Consistent gene tree G’ that satisfies the relations of R such that…

…the Robinson-Foulds distance between G and G’ is minimized.

Problem 4

Given : a species tree S, a gene tree G, and a partial set R of orthologs and paralogs to satisfy.

Find : an S-Consistent gene tree G' that satisfies the relations of R such that…

…the Robinson-Foulds distance between G and G' is minimized.

If R has only orthologs, or only paralogs, then this is feasible in polynomial time (Lafond & al. 2013).

Problem 4

Given : a species tree S, a gene tree G, and a partial set R of orthologs and paralogs to satisfy.

Find : an S-Consistent gene tree G' that satisfies the relations of R such that…

…the Robinson-Foulds distance between G and G' is minimized.

If R has both orthologs and paralogs, then…

Problem 4

Given : a species tree S, a gene tree G, and a partial set R of orthologs and paralogs to satisfy.

Find : an S-Consistent gene tree G' that satisfies the relations of R such that…

…the Robinson-Foulds distance between G and G' is minimized.

If R has both orthologs and paralogs, then…

Surprise !

NP-COMPLETE (Lafond and El-Mabrouk, 2015)

Reduction from the gene removal problem (Problem 2)

More problems

The inferred relations could be **weighted**, e.g. associated with a **probability** of orthology.

A more general problem : edge insertion / deletion where each operation has a **distinct cost**.

More problems

The inferred relations could be **weighted**, e.g. associated with a **probability** of orthology.

A more general problem : edge insertion / deletion where each operation has a **distinct cost**.

- Generalizes both **Problem 1** (edge insertion / deletion) and **Problem 3** (gene tree correction, minimize broken relations).

More problems

The inferred relations could be **weighted**, e.g. associated with a **probability** of orthology.

A more general problem : edge insertion / deletion where each operation has a **distinct cost**.

- Generalizes both **Problem 1** (edge insertion / deletion) and **Problem 3** (gene tree correction, minimize broken relations).
- Factor O(n) approximation algorithm (unpublished). We do not know whether it is essentially best possible.

Conclusion

- Gene tree => Orthologs / paralogs : **YES**
- Orthologs / paralogs => Gene tree : **NOT ALWAYS**

- Unfortunately, error correction of orthology / paralogy relations is difficult.

- Even assuming relations are error-free, gene tree coorrection is difficult.