ORTHOLOGY AND PARALOGY CONSTRAINTS: SATISFIABILITY AND CONSISTENCY

Manuel Lafond, Nadia El-Mabrouk

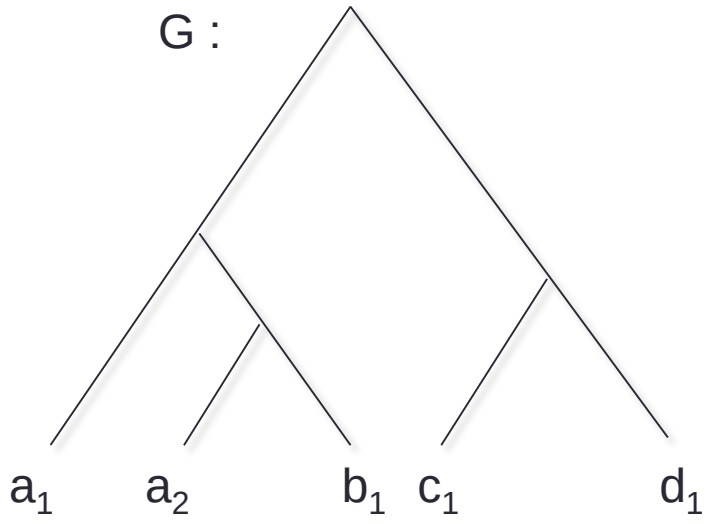
University of Montreal

Outline

- **Introduction**
  - Gene trees, orthologs, paralogs, …
- **3 problems, given a set of orthologs and paralogs**
  - Satisfiability
  - Consistency with a species tree S
  - Self-consistency
- **Experiments**

Introduction

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

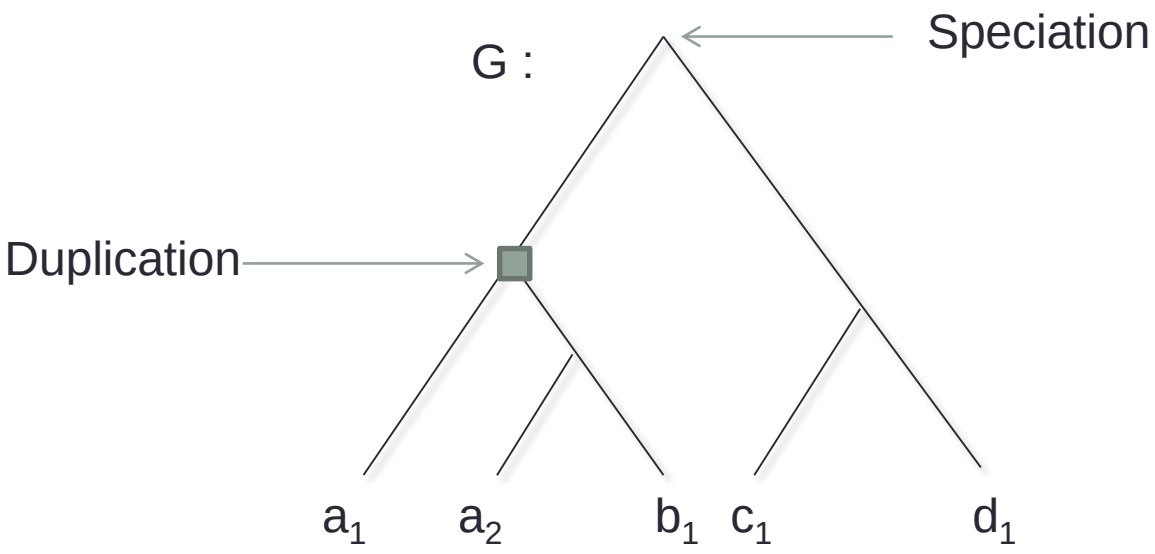


a,b,c,d are species

Gene trees don't have to be binary.

Introduction

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



Introduction

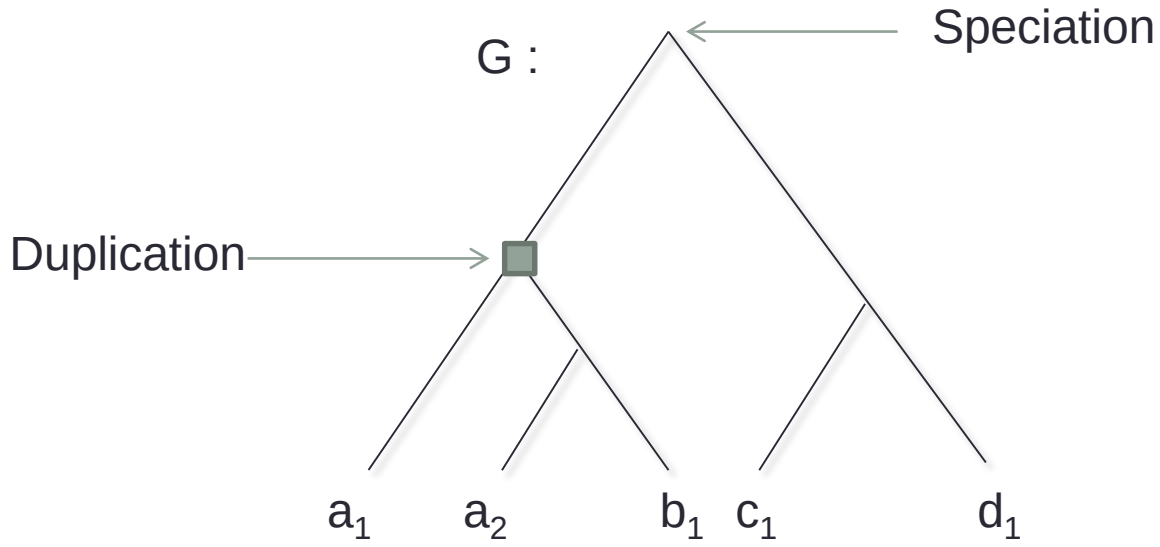
Orthologs : LCA has undergone speciation

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

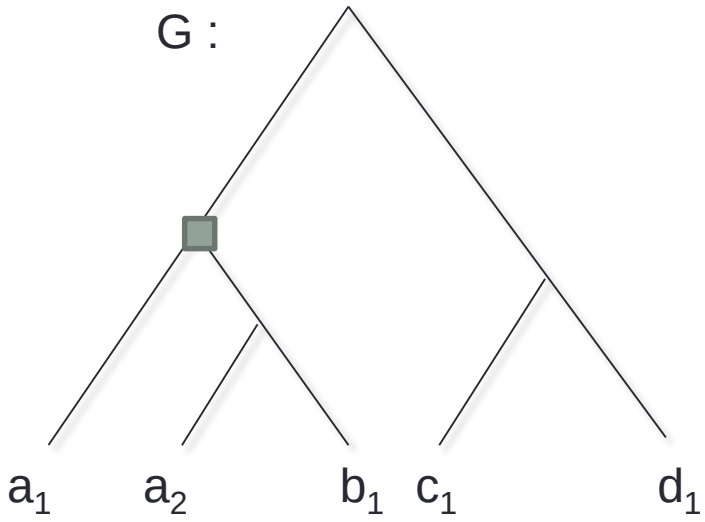


Introduction

If we have G (and trust its Dup/Spec labeling), then we have **all orthology/paralogy** relationships.

Orthologs

a_1b_1
a_1c_1
a_1d_1
a_2c_1
a_2d_1
b_1c_1
b_1d_1
c_1d_1



Paralogs

a_1a_2
a_1b_1

Introduction

How does that go the **other way around** ?

If we have the orthology/paralogy relationships, can we get the gene tree ?

Orthologs

a_1b_1
a_1c_1
a_1d_1
a_2c_1
a_2d_1
b_1c_1
b_1d_1
c_1d_1

?

Paralogs

a_1a_2
a_1b_1

Introduction

Various software let us infer orthology (and sometimes paralogy) without a gene tree

Sequence-based

COG	(Tatusov, Galperin, Natale & Koonin, 2000)
OrthoMCL	(Li, Stoeckert & Roos, 2003)
InParanoid	(Berglund, Sjolund, Ostlund & Sonnhammer, 2008)
Proteinortho	(Findeib, Steiner, Marz, Stadler & Prohaska, 2011)
…	

Gene order-based

GIGA	(Thomas, 2010)
SYNERGY	(Wapinski, Pfeffer, Friedman & Regev, 2007)
[Unnamed]	(Lafond, Swenson, El-Mabrouk, 2013)

Introduction

Various software let us infer orthology (and sometimes paralogy) without a gene tree

Sequence-based

- COG
- OrthoMCL
- InParanoid
- Proteinortho
- …

Gene order-based

- GIGA
- SYNERGY
- [Unnamed]

None of them finds ALL orthologies/paralogies !

Satisfiability

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

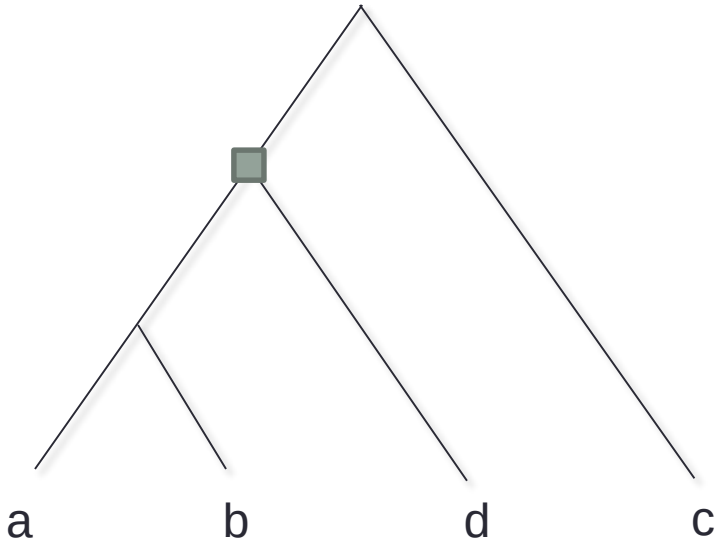
Paralogs = (a, d) (b, d)

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

Satisfiability

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

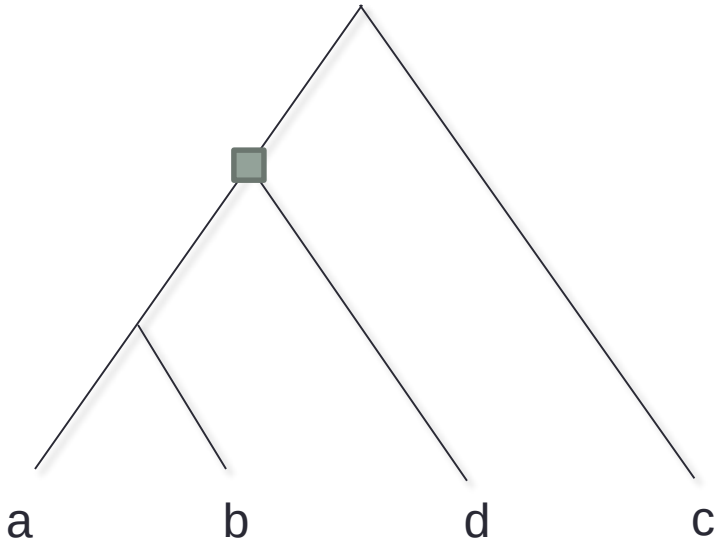
Paralogs = (a, d) (b, d)



Satisfiability

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

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



Satisfiability

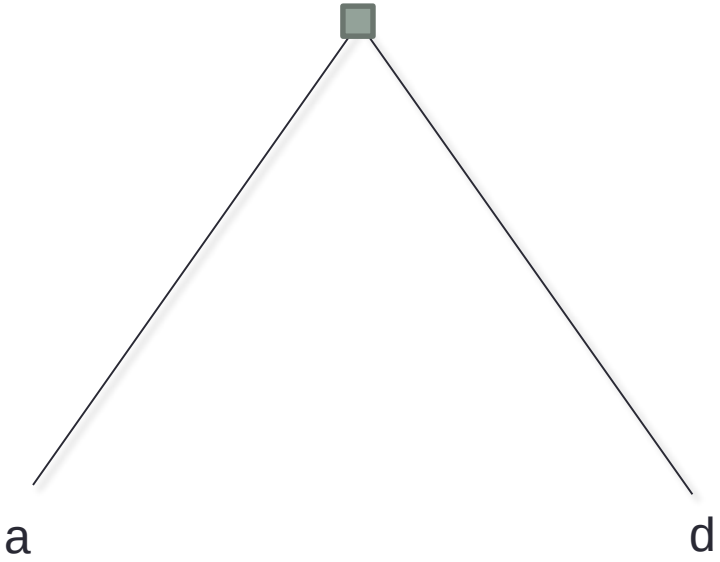
Orthologs = (a,b) (a, c) (c, d)

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

Satisfiability

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

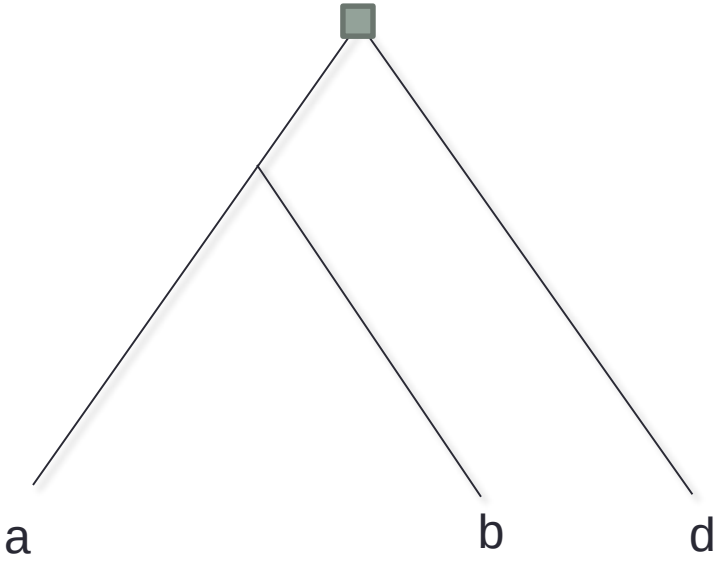
Paralogs = (a, d) (b, c) (b, d)



Satisfiability

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

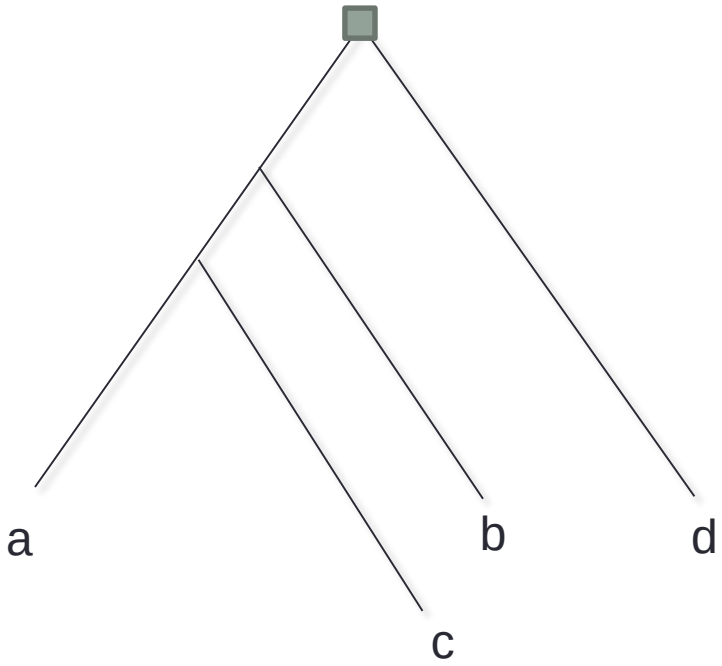
Paralogs = (a, d) (b, c) (b, d)



Satisfiability

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

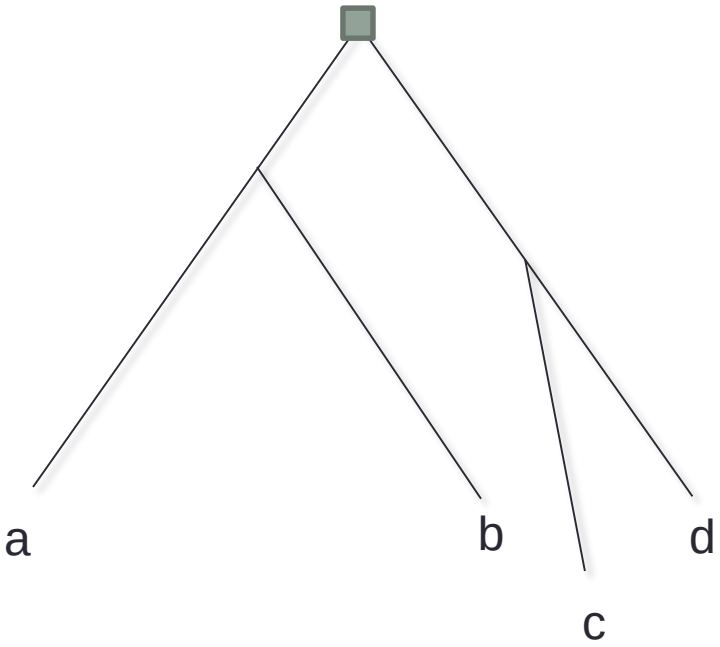
Paralogs = (a, d) (b, c) (b, d)



Satisfiability

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

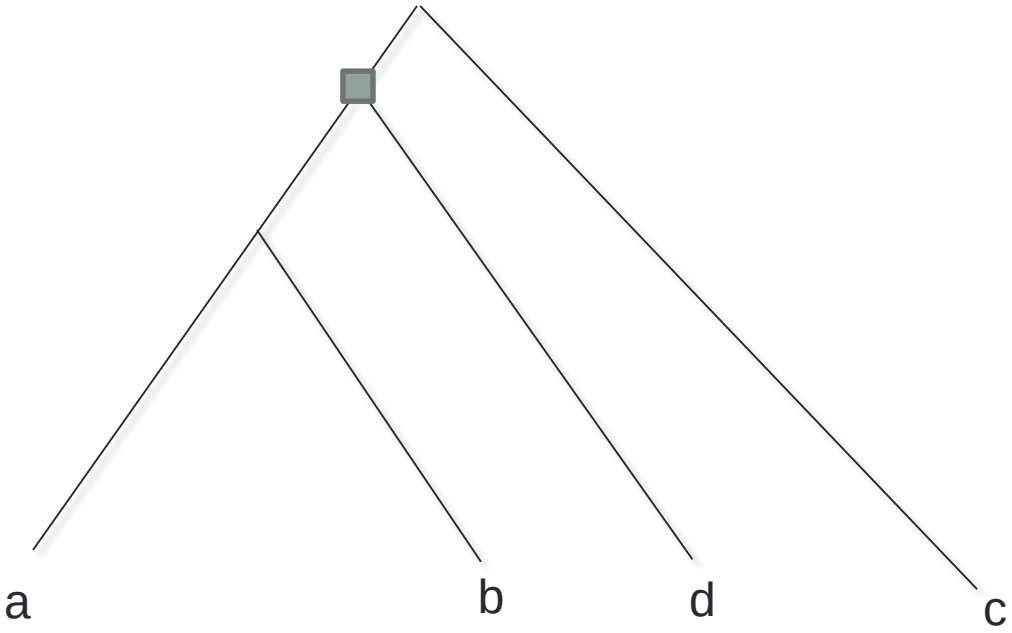
Paralogs = (a, d) (b, c) (b, d)



Satisfiability

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

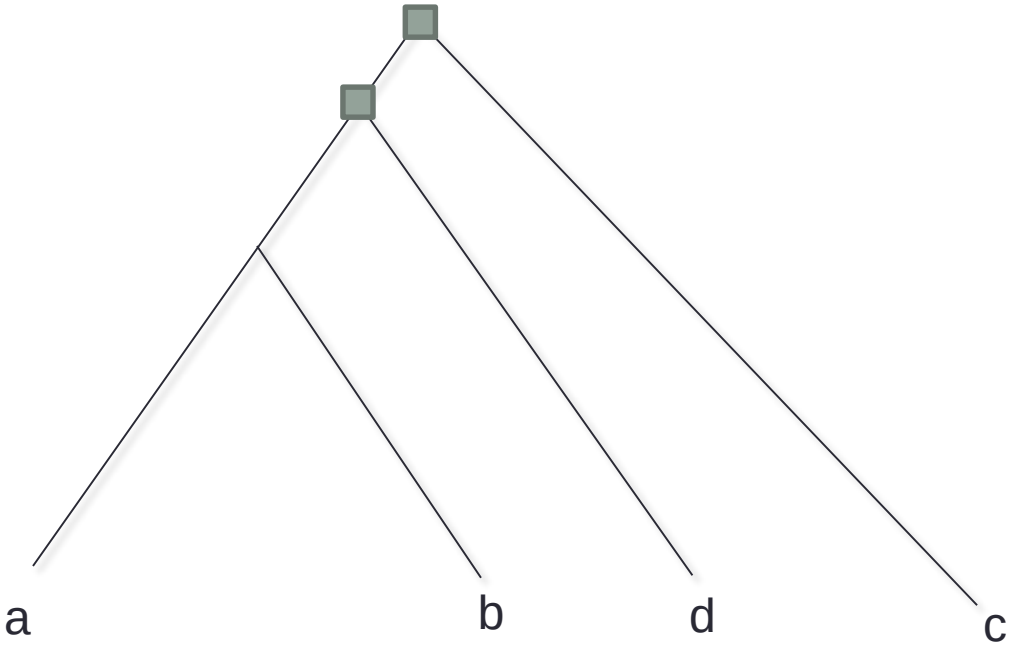
Paralogs = (a, d) (b, c) (b, d)



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 !

Consistency with a species tree S

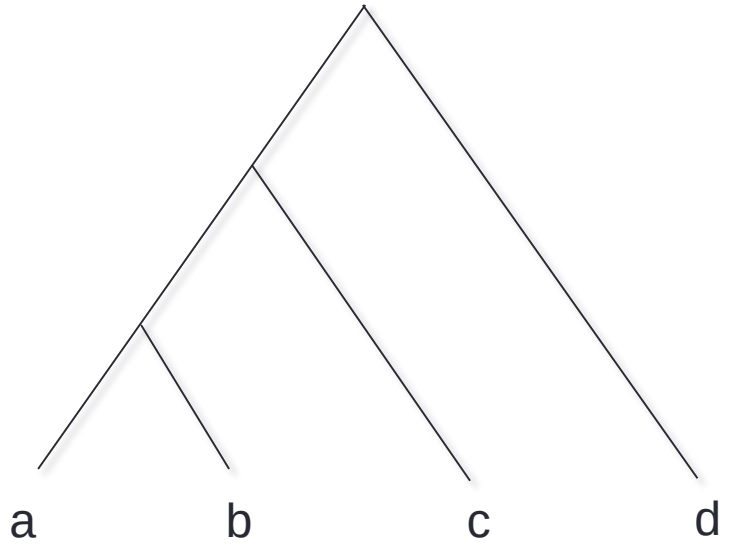
Orthologs = (a,d) (c,d)

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

Gene tree G

?

Species tree S

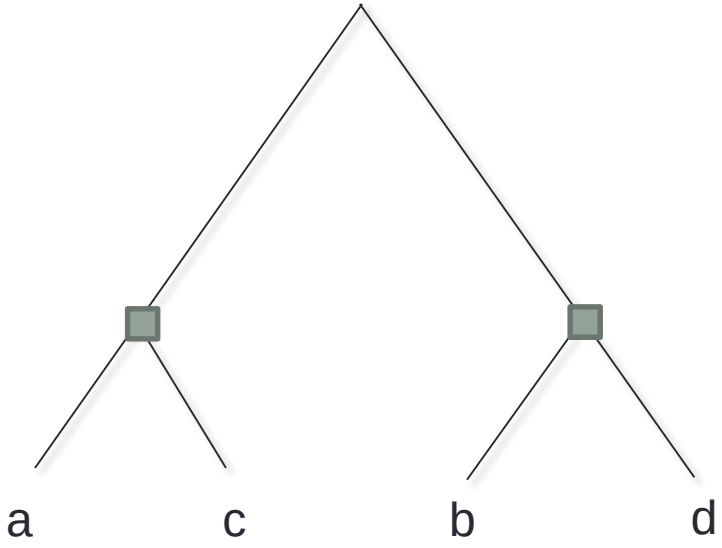


Consistency with a species tree S

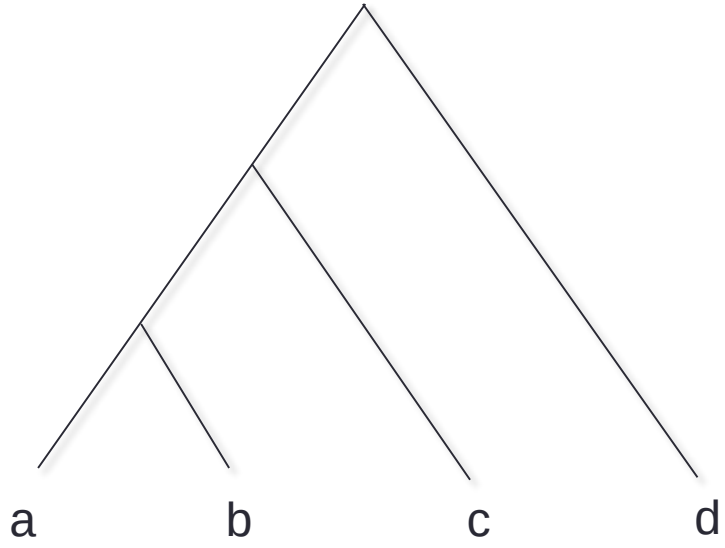
Orthologs = (a,d) (c,d)

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

Gene tree G

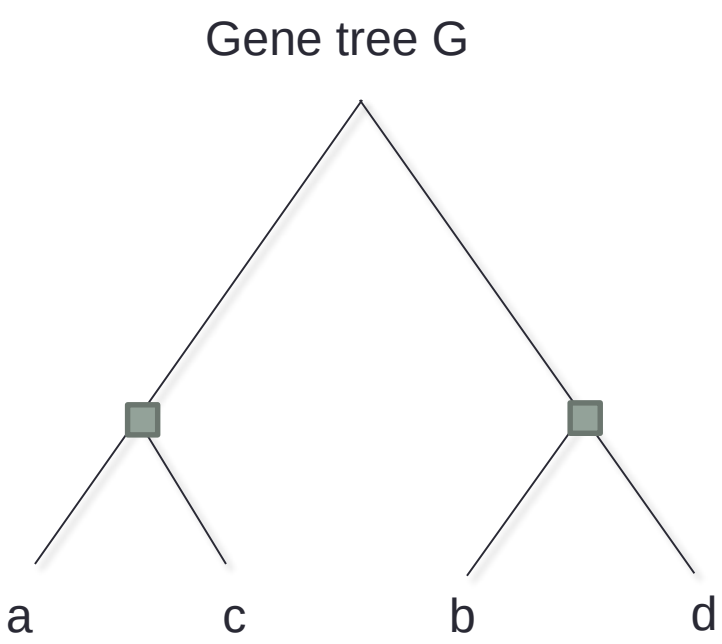


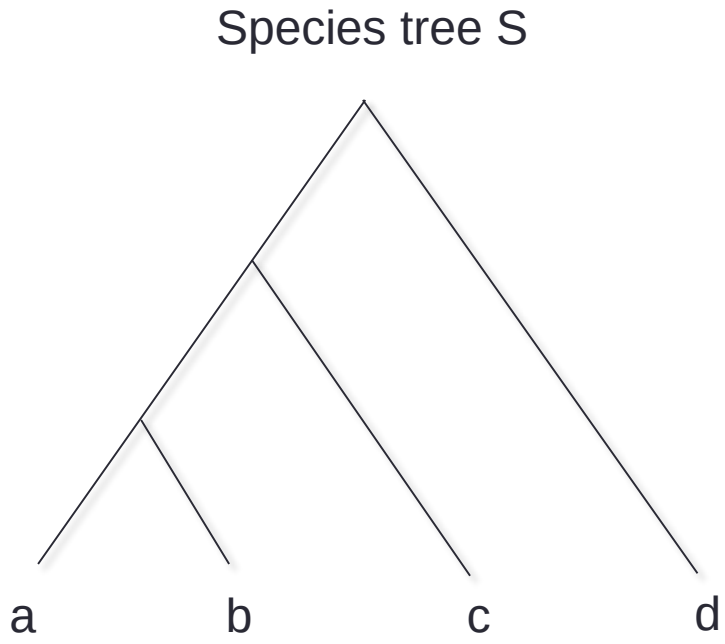
Species tree S



Consistency with a species tree S

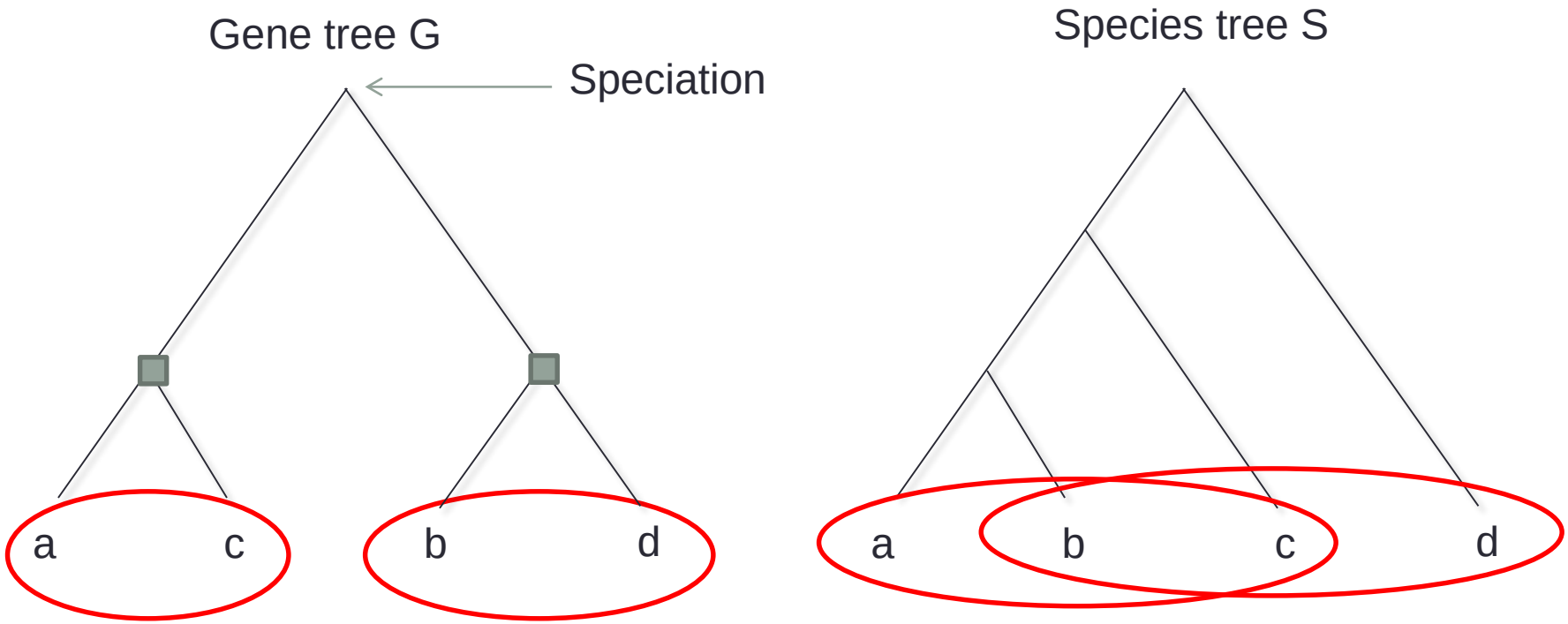
Consistency with a species tree S : If genes from species sets X,Y are separated by speciation in G, then species X, Y are separated in S.





Consistency with a species tree S

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



Consistency with a species tree S

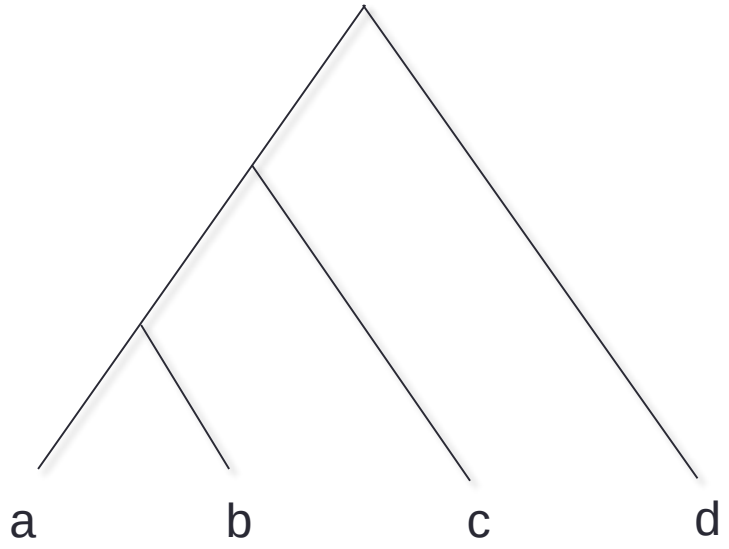
Orthologs = (a,d) (c,d)

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

Gene tree G

?

Species tree S

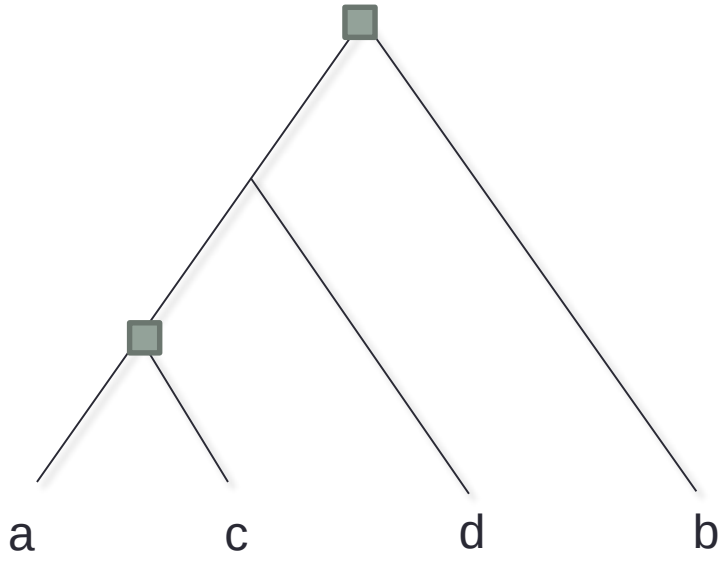


Consistency with a species tree S

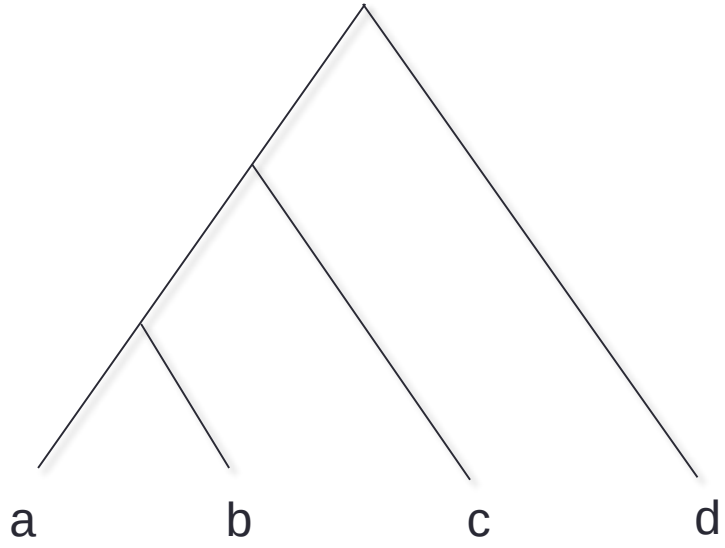
Orthologs = (a,d) (c,d)

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

Gene tree G



Species tree S

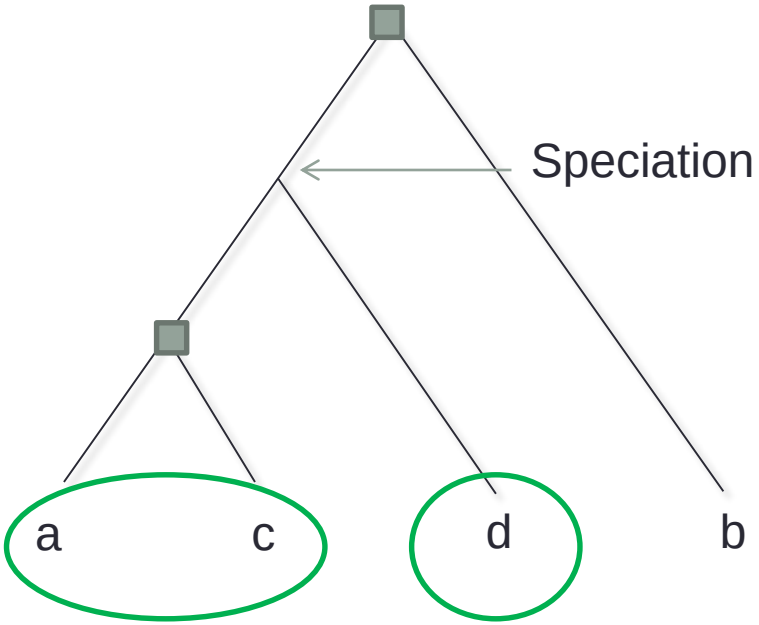


Consistency with a species tree S

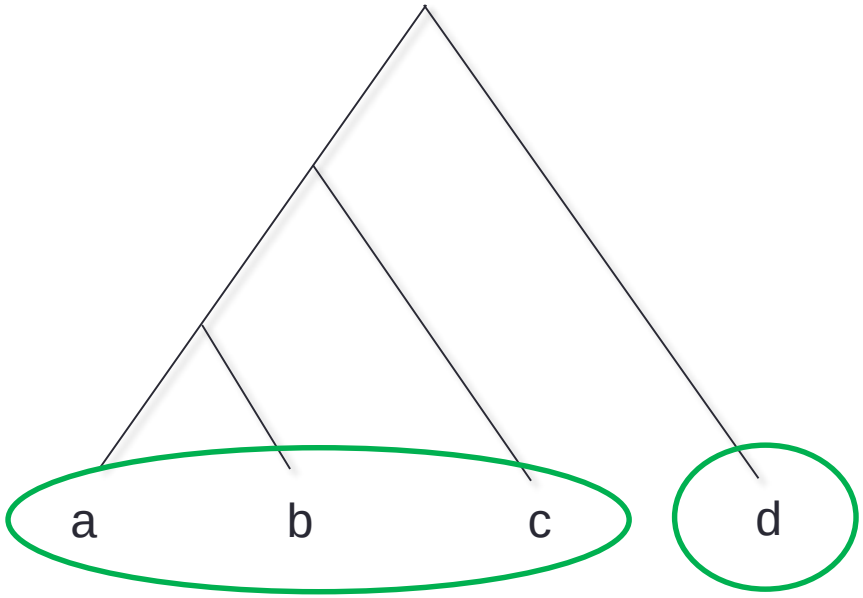
Orthologs = (a,d) (c,d)

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

Gene tree G



Species tree S



Self-consistency

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

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

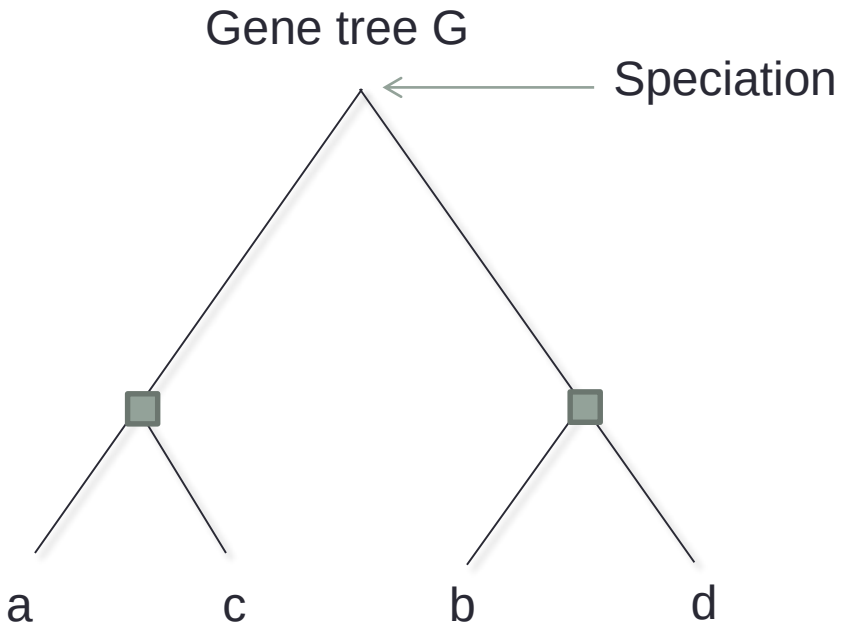
Can we build a gene tree G displaying these relationships such that

there exists some species tree S consistent with it ?

Self-consistency

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

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



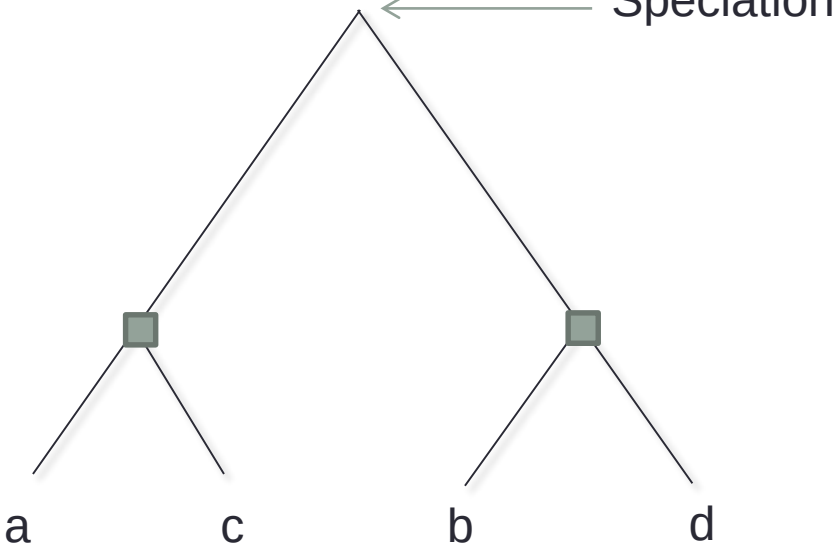
Self-consistency

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

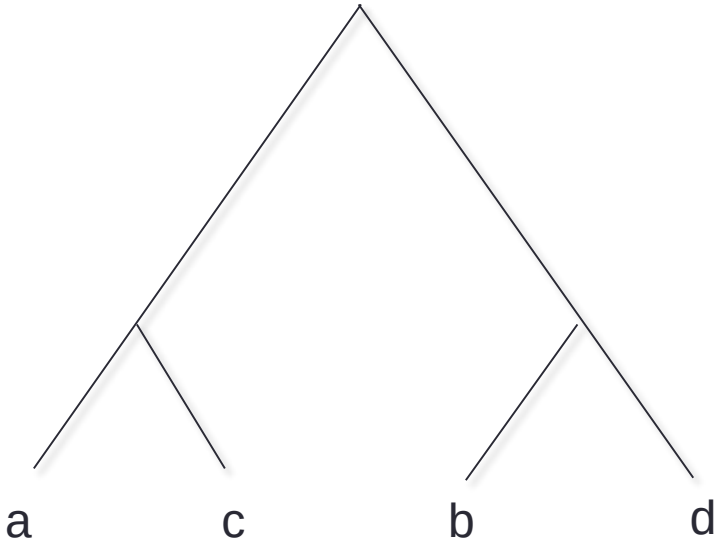
Paralogs = (a,c) (b, d)

Gene tree G

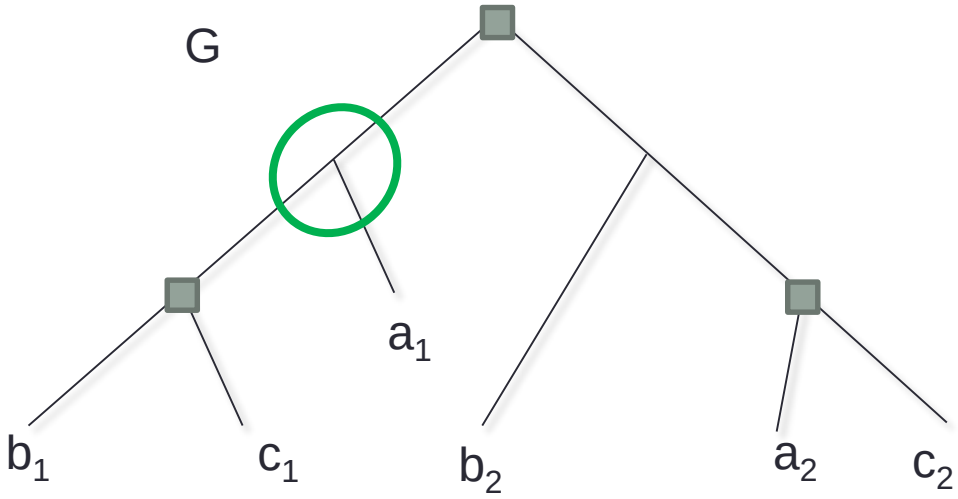
Speciation

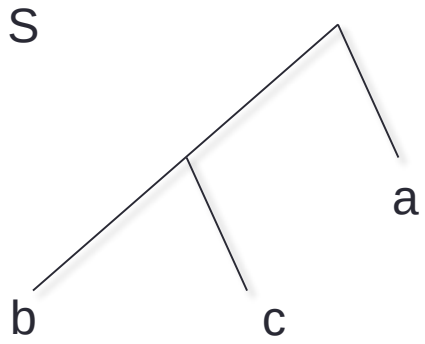


Species tree S

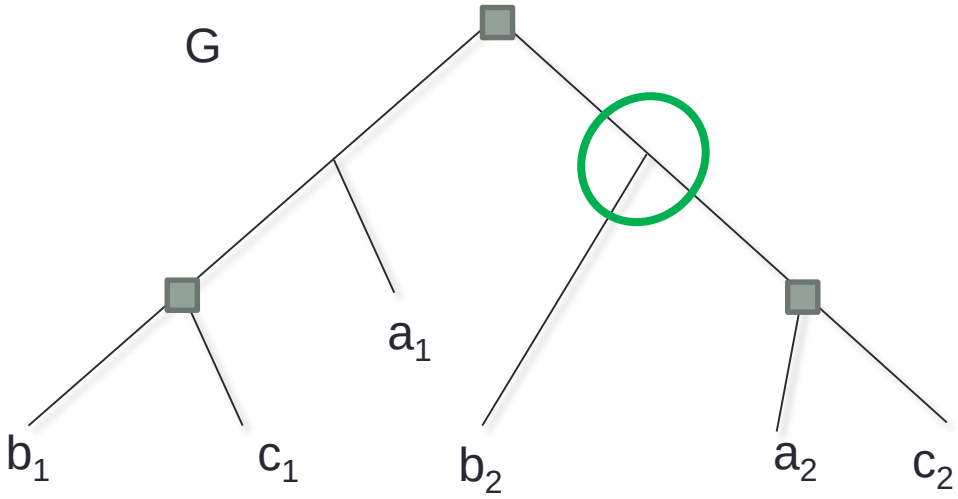


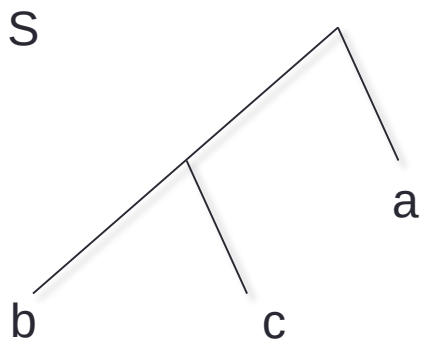
Not self-consistent

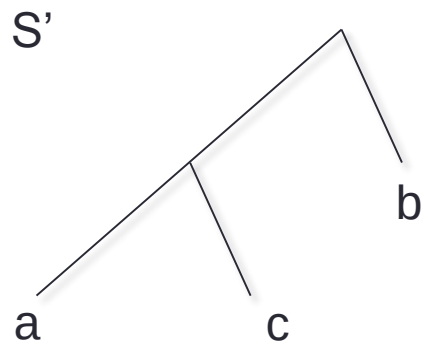




Not self-consistent







The problem(s)

Given a set C of orthologs and paralogs :

1. Is C **satisfiable** ?

   Does there exist a DS-tree that exhibits all relationships in C ?
2. Is C **consistent with a given species tree S** ?

   Is there some DS-tree that satisfies C that is also consistent with S ?
3. Is C **self-consistent** ?

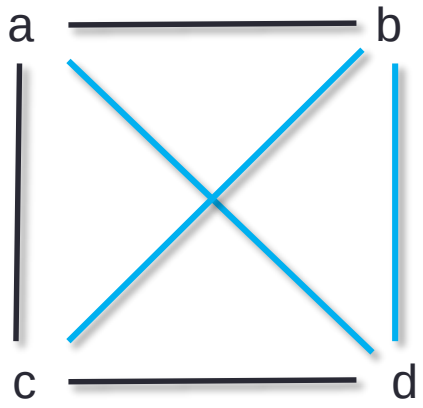
   Is there some species tree that C is consistent with ?

Satisfiability

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

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

Constraint graph R



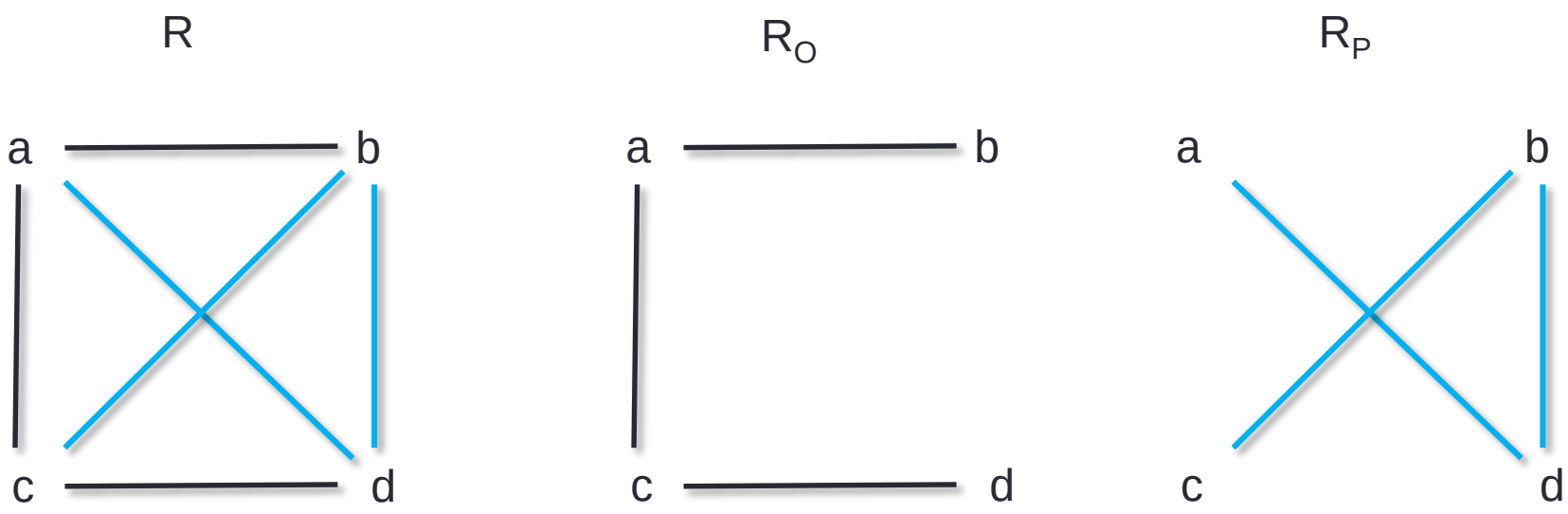
Orthologs

Paralogs

Satisfiability

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

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



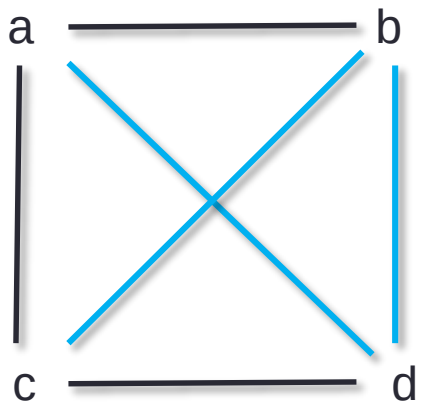
Satisfiability

(Hernandez-Rosales & al., 2012)

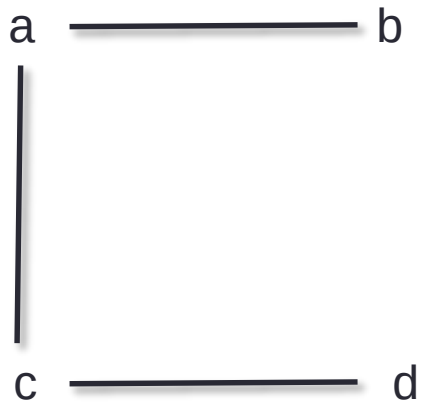
If R is a complete graph, then the given set of relationships is satisfiable iff

R_O is P_4-free (and equivalently, if R_P is P_4-free)

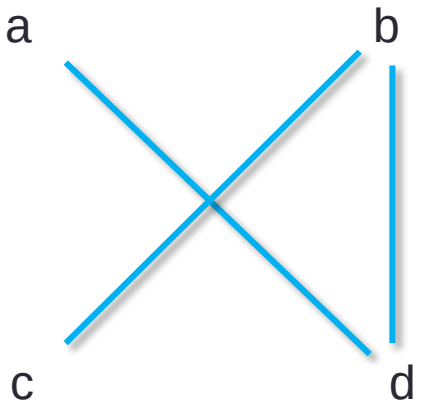
R



R_O



R_P



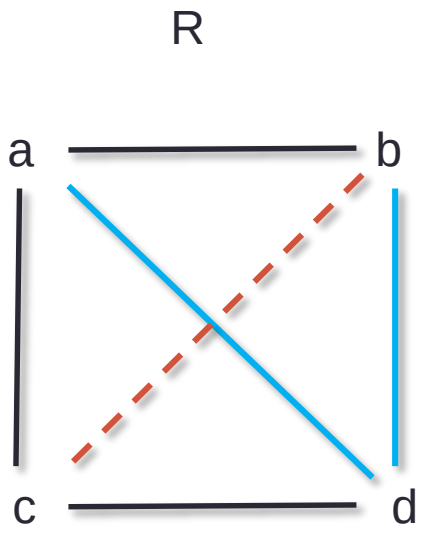
Orthologs

Paralogs

Unknown relationships

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

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

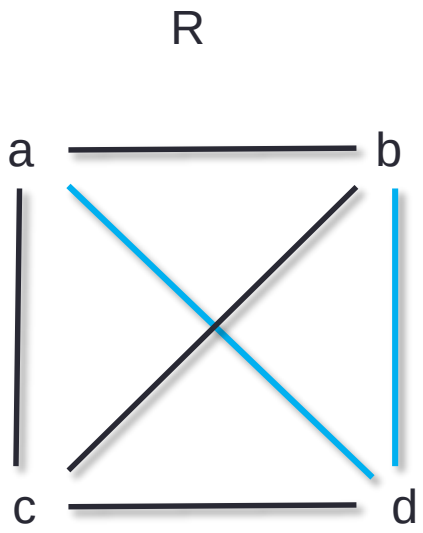


The (b,c) relationship is **unknown**.
Our relationships are satisfiable iff we can decide the (b,c) relationship such that RO will be P_4-free

Unknown relationships

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

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

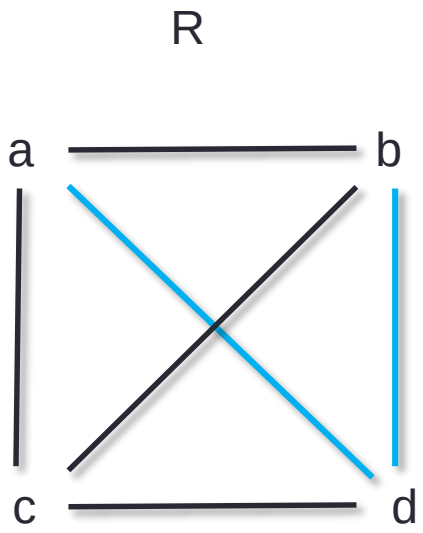


The (b,c) relationship is **unknown**.
Our relationships are satisfiable iff we can decide the (b,c) relationship such that RO will be P_4-free

Unknown relationships

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

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



The (b,c) relationship is **unknown**.
Our relationships are satisfiable iff we can decide the (b,c) relationship such that RO will be P_4-free

This problem is equivalent to the **Graph Sandwich Problem** on the class of cographs

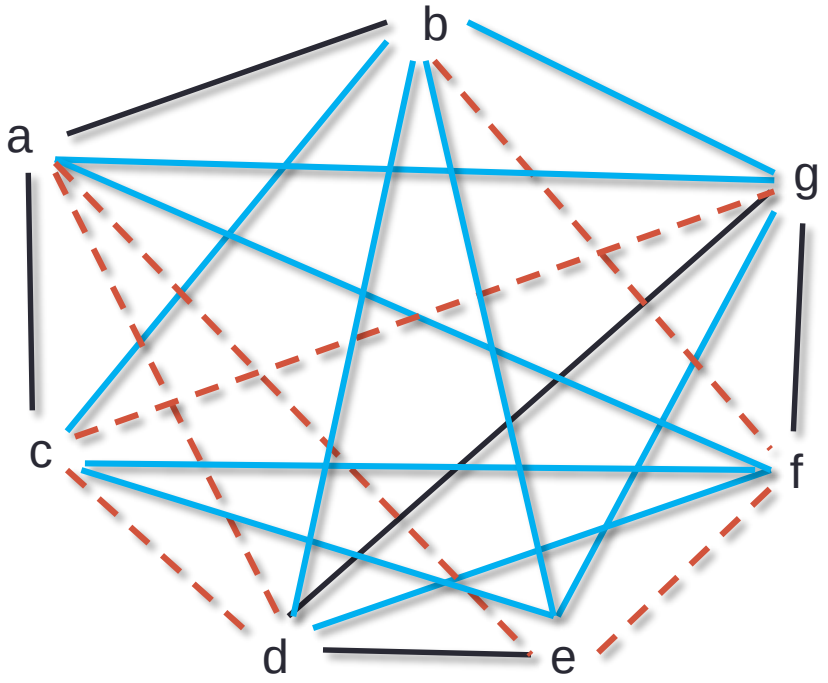
Satisfiability

Theorem (Golumbic, Kaplan and Shamir, 1994) :

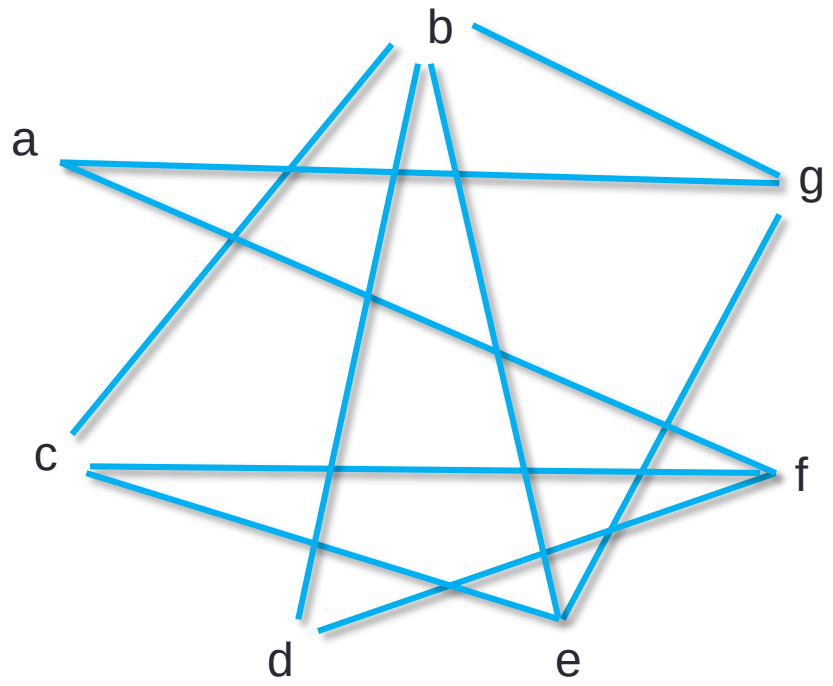
A relationship graph R is satisfiable iff **at least** one of the following holds :

1) R_O is disconnected, and each of its component is satisfiable
2) R_P is disconnected, and each of its component is satisfiable

Constructing a gene tree

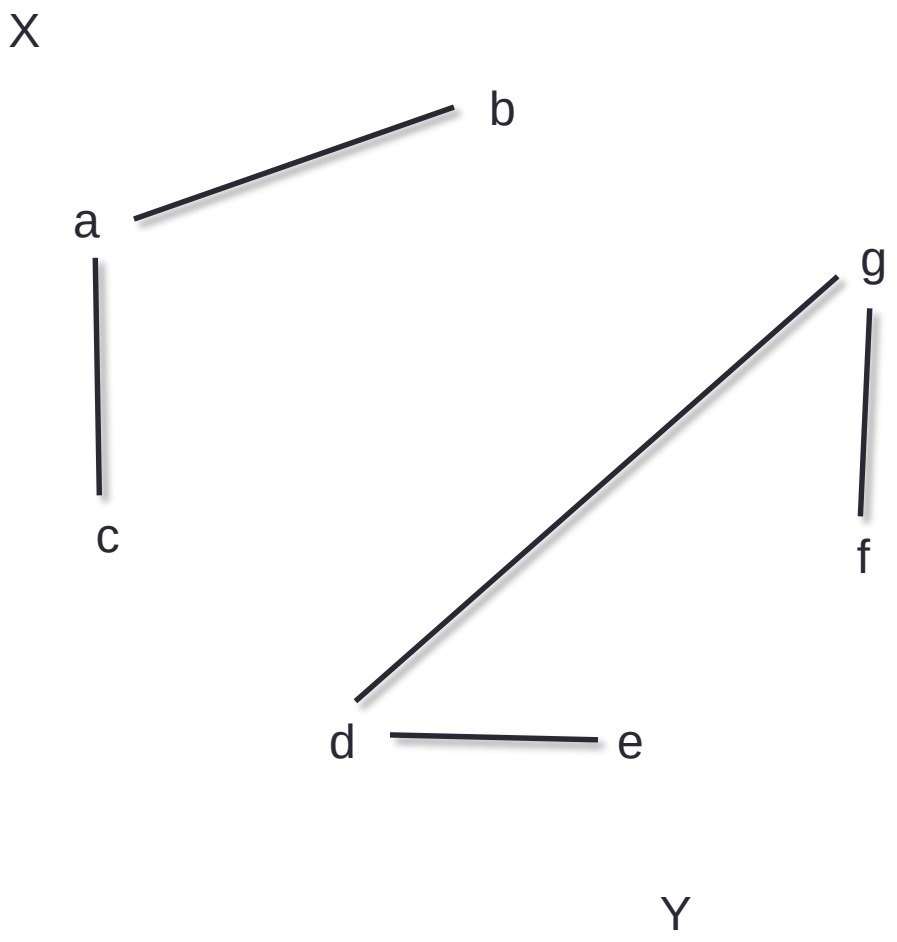


Constructing a gene tree



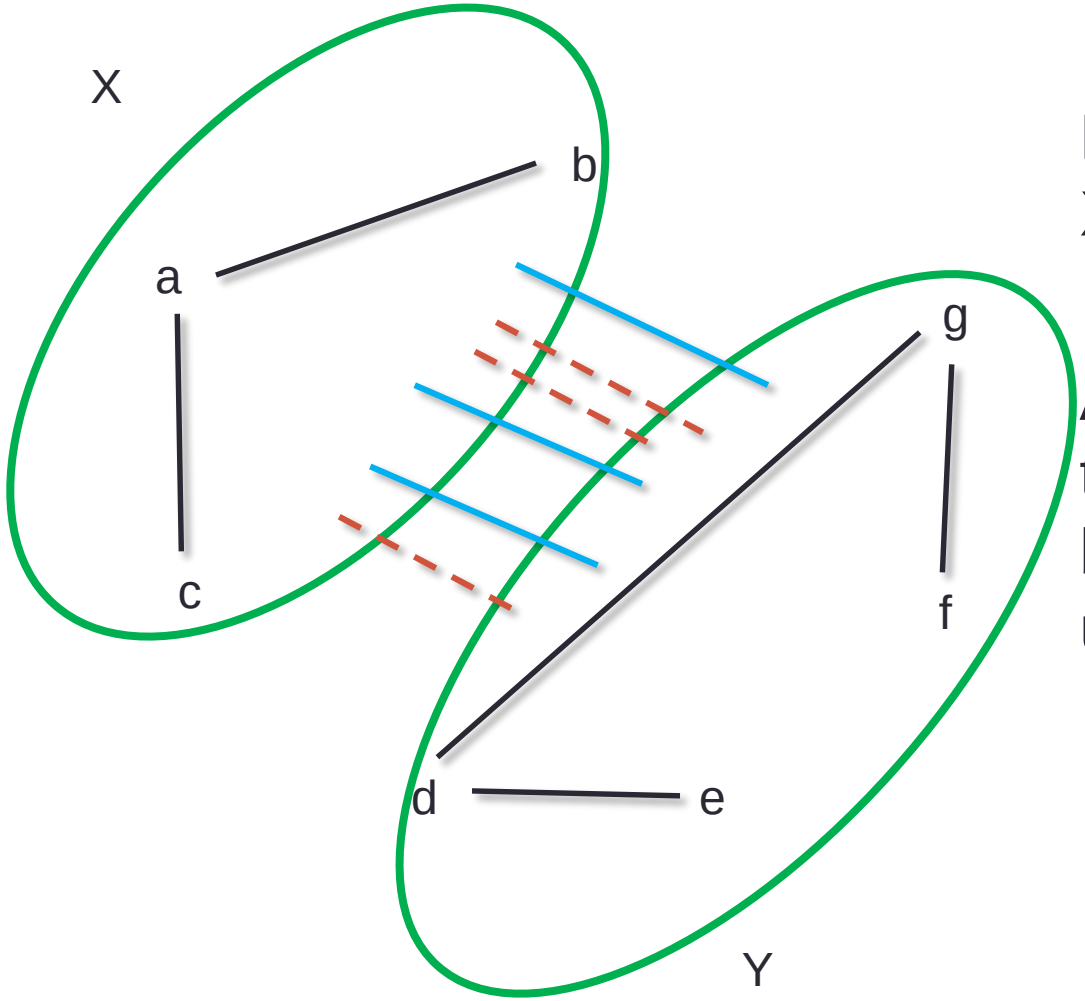
R_P is connected, nothing to do here.

Constructing a gene tree



R_O has 2 components, X and Y.

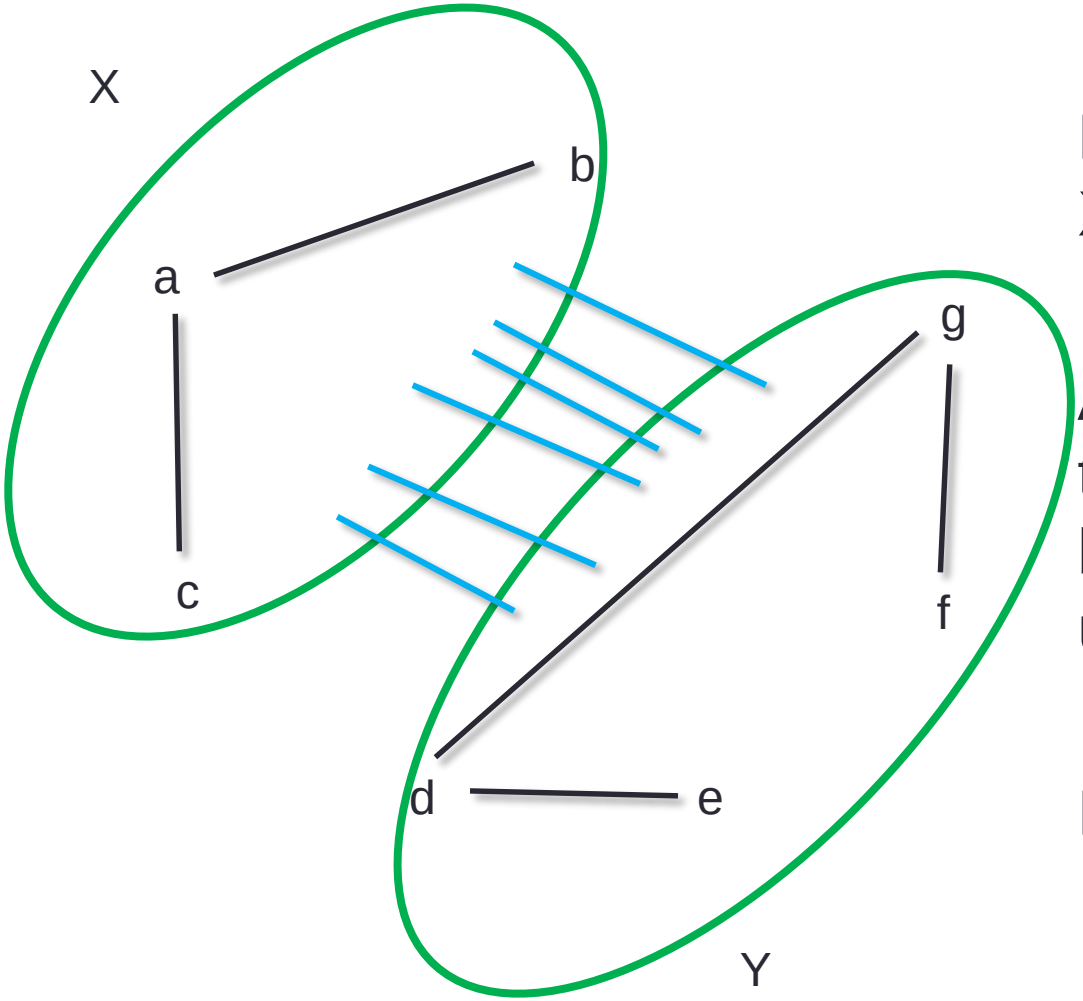
Constructing a gene tree



R_O has 2 components, X and Y.

All edges going from X to Y are either black or blue (paralogy or unknown).

Constructing a gene tree

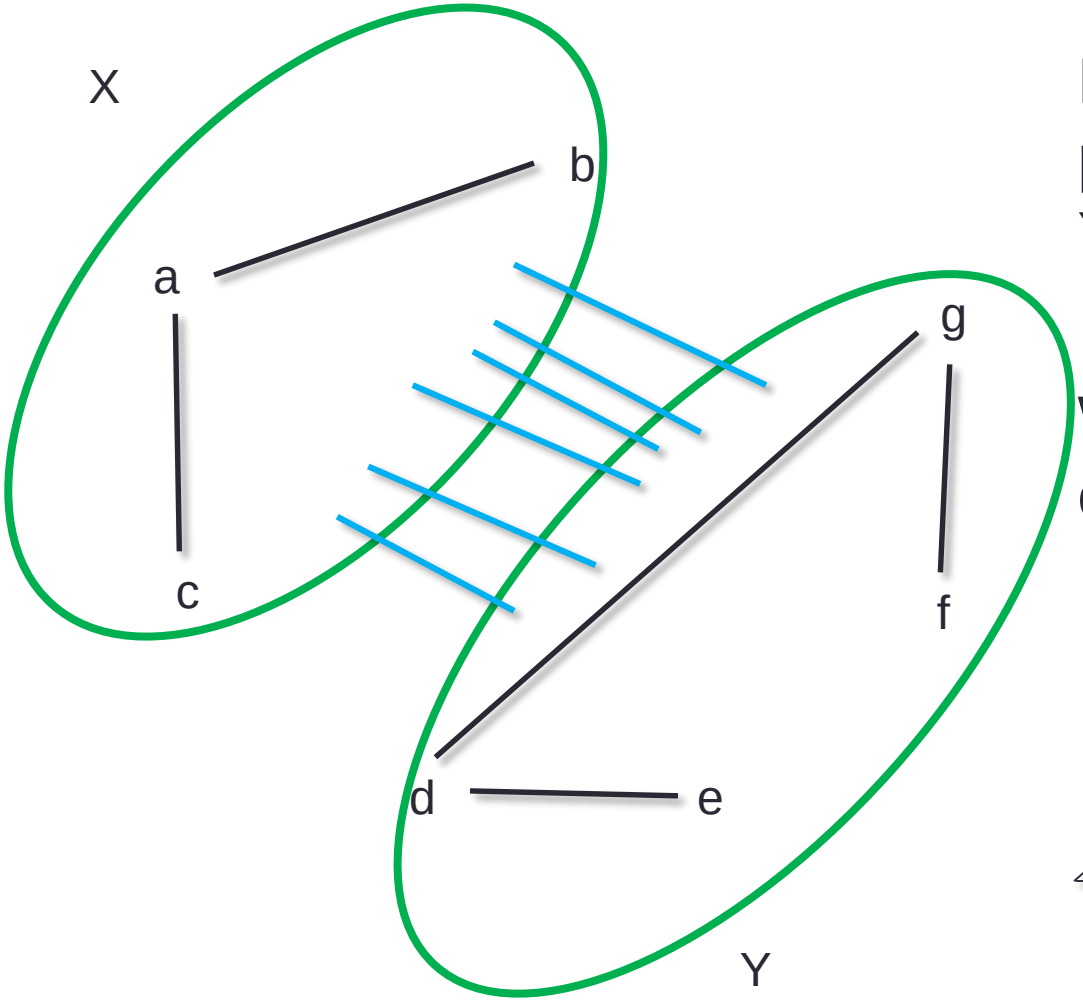


R_O has 2 components, X and Y.

All edges going from X to Y are either black or blue (paralogy or unknown).

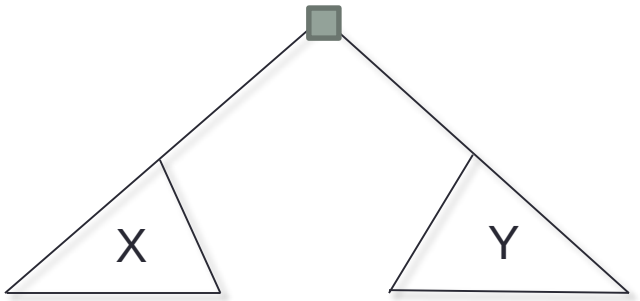
Make it all blue !

Constructing a gene tree

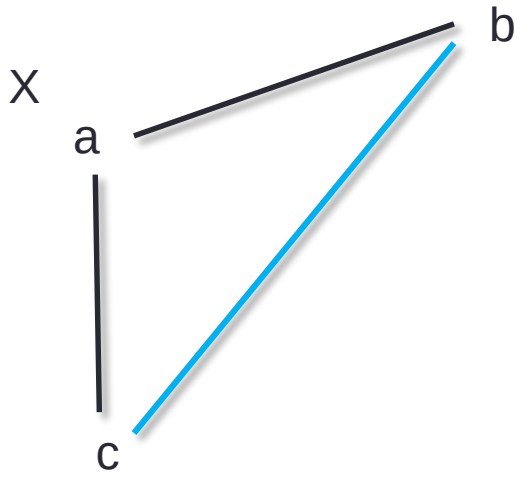


Now, all genes of X are paralog to all genes of Y.

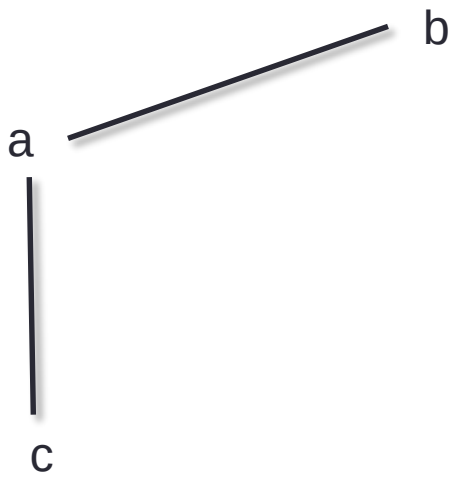
We can start building our gene tree as such :



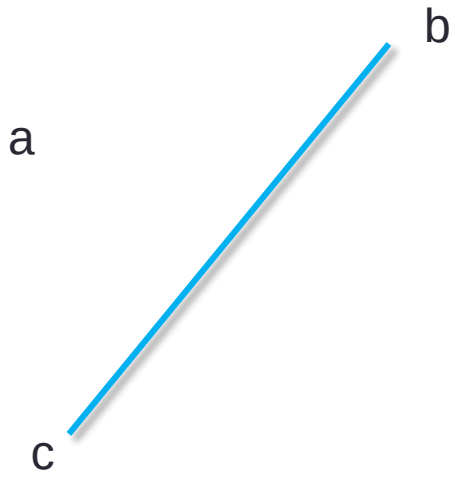
Constructing a gene tree



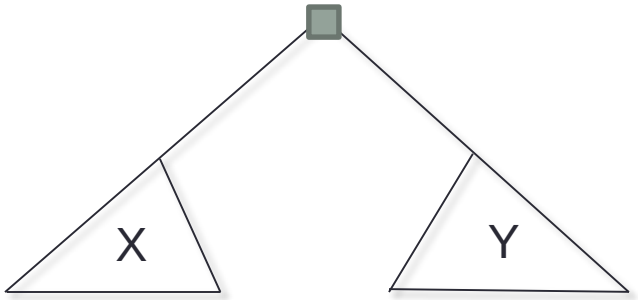
Repeat with X, and Y.



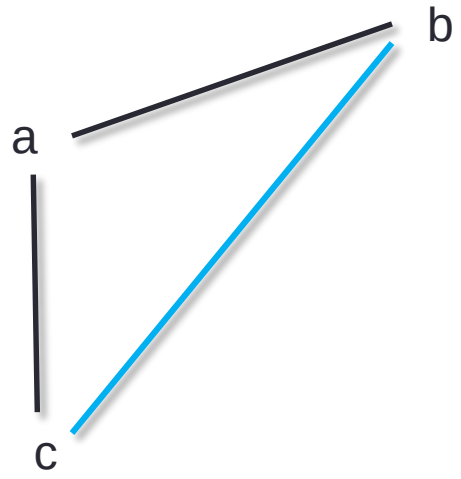
$R_O[X]$



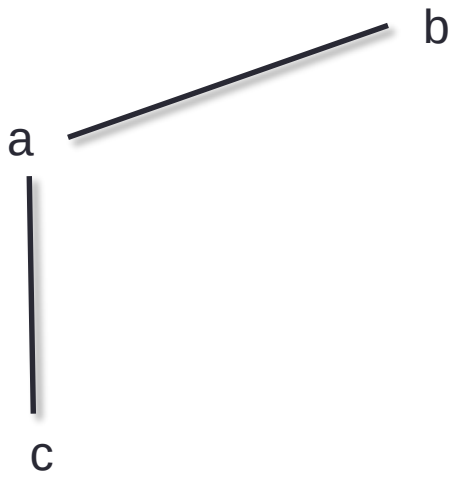
$R_P[X]$



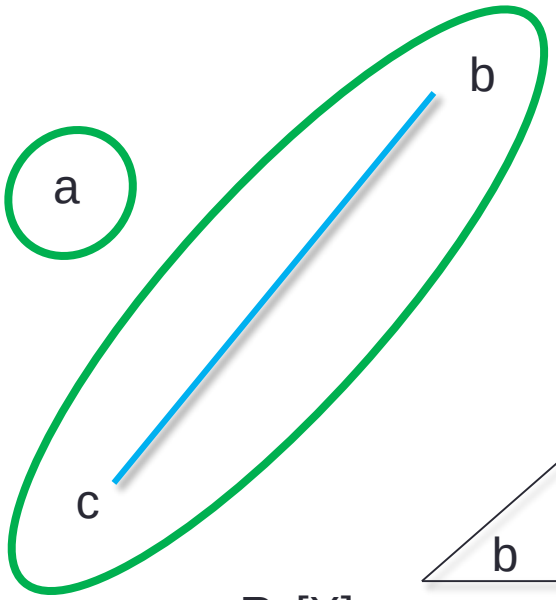
Constructing a gene tree



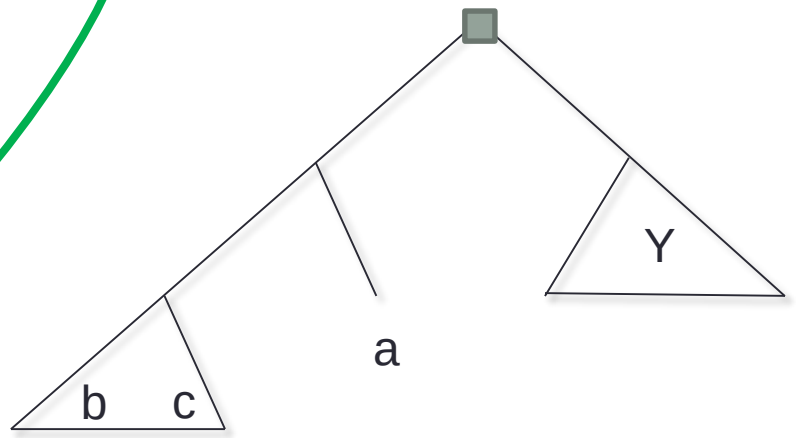
Repeat with X, and Y,



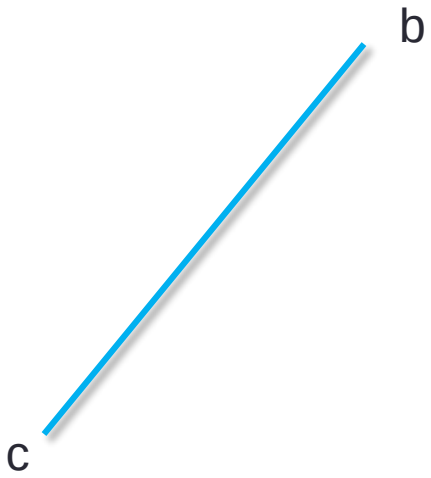
$R_O[X]$



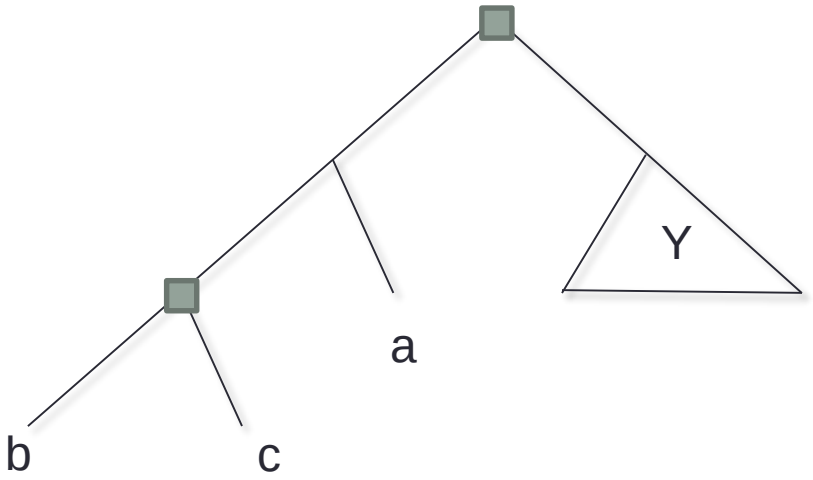
$R_P[X]$



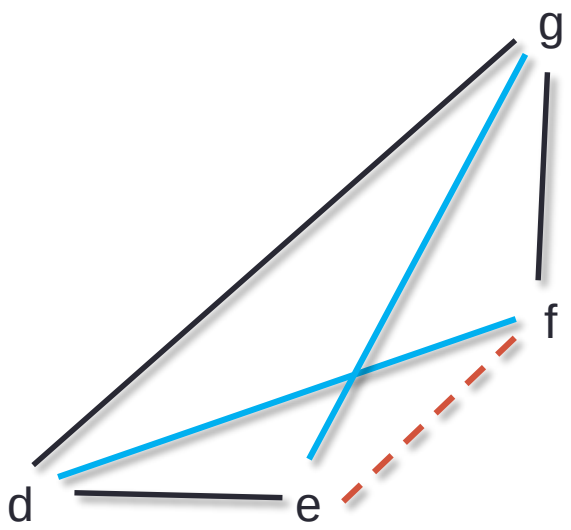
Constructing a gene tree



Repeat with X, and Y.

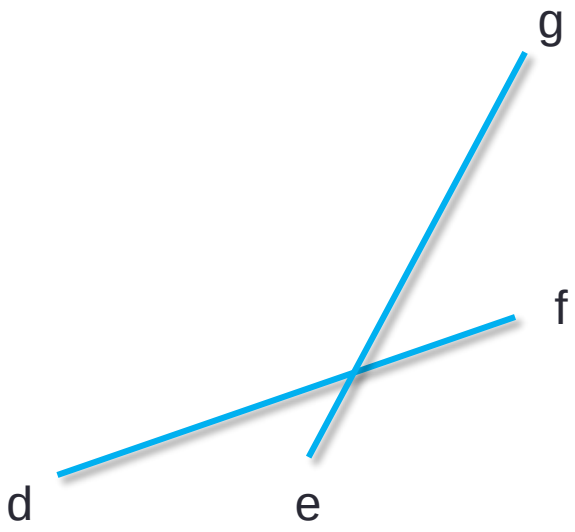


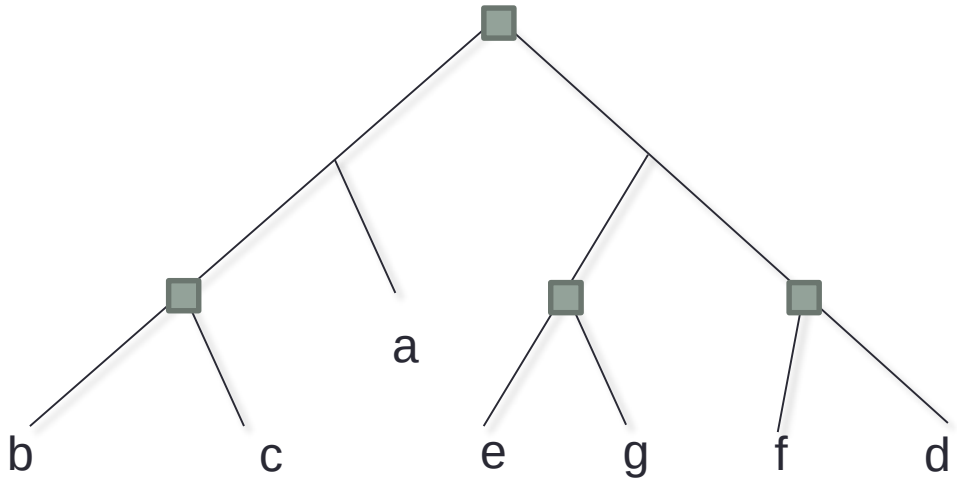
Constructing a gene tree



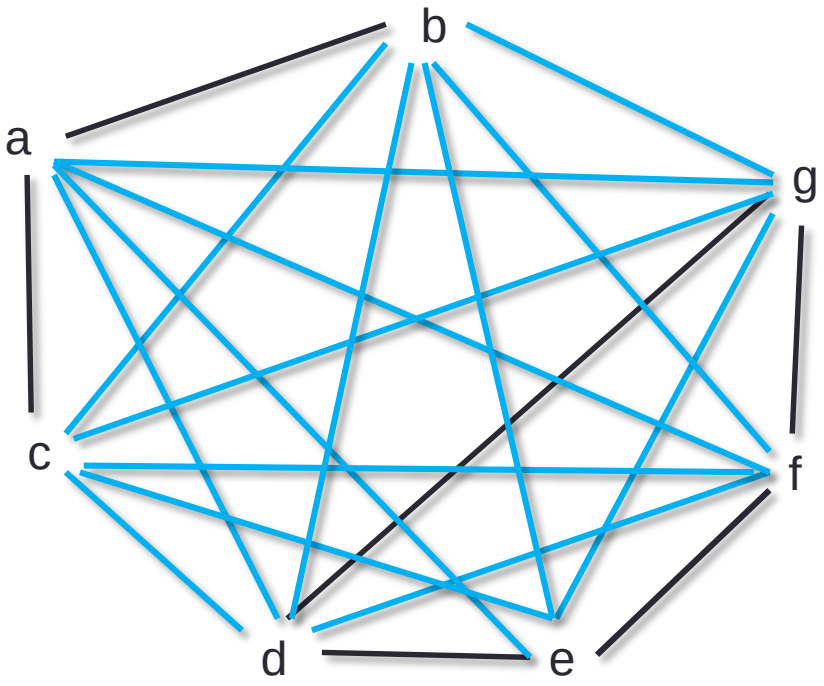
Repeat with X, and Y.

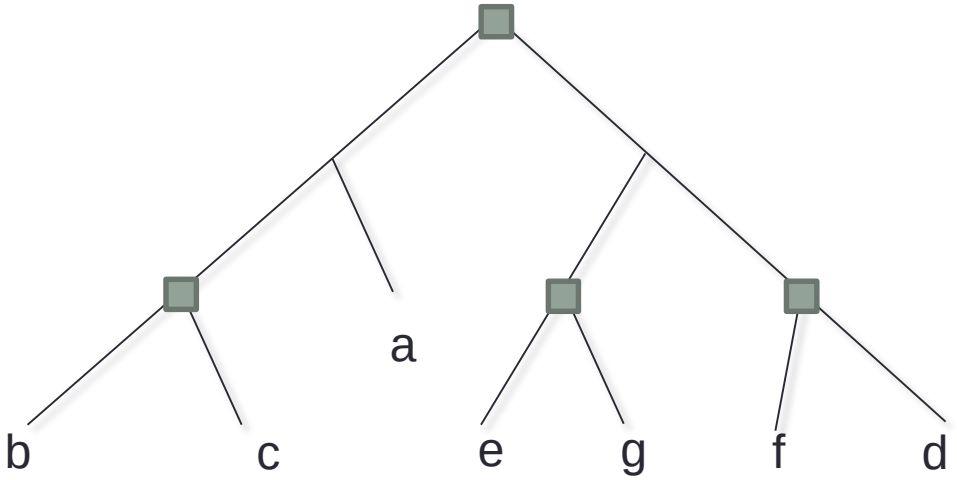
$R_P[Y]$



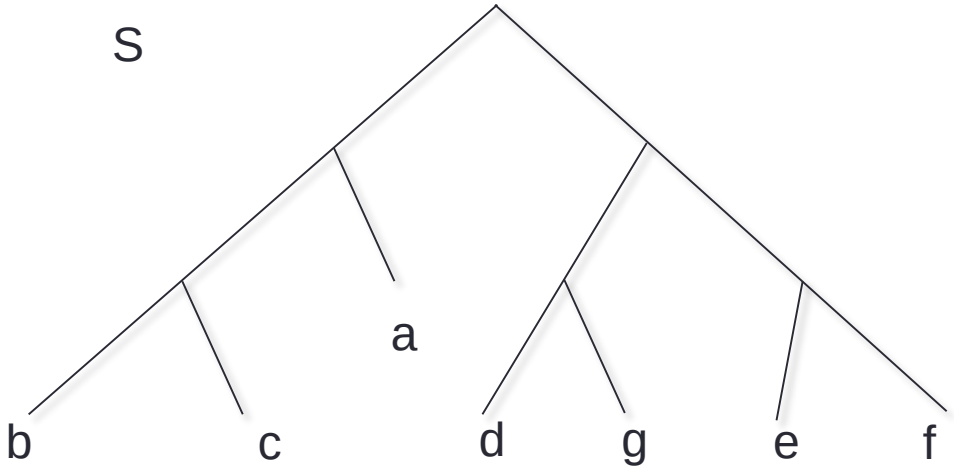


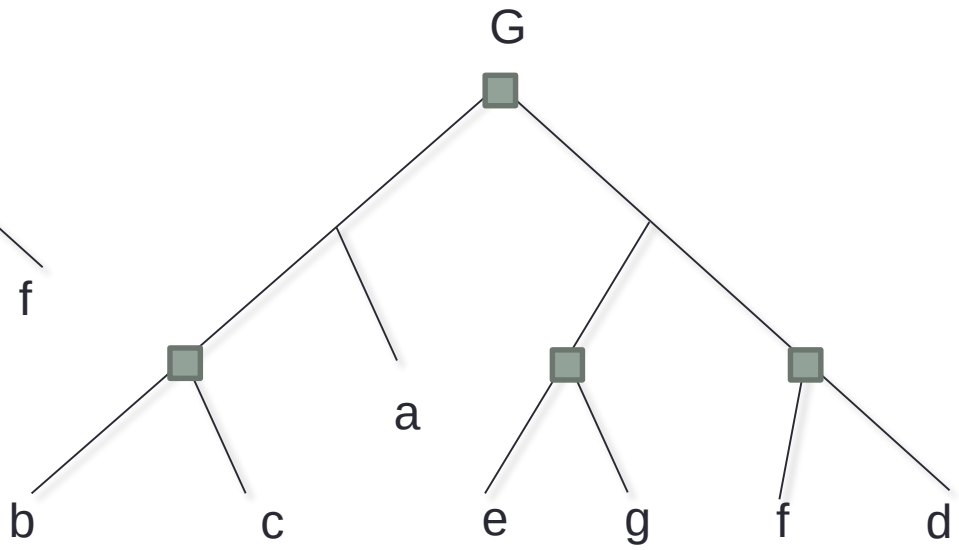
Constructing a gene tree





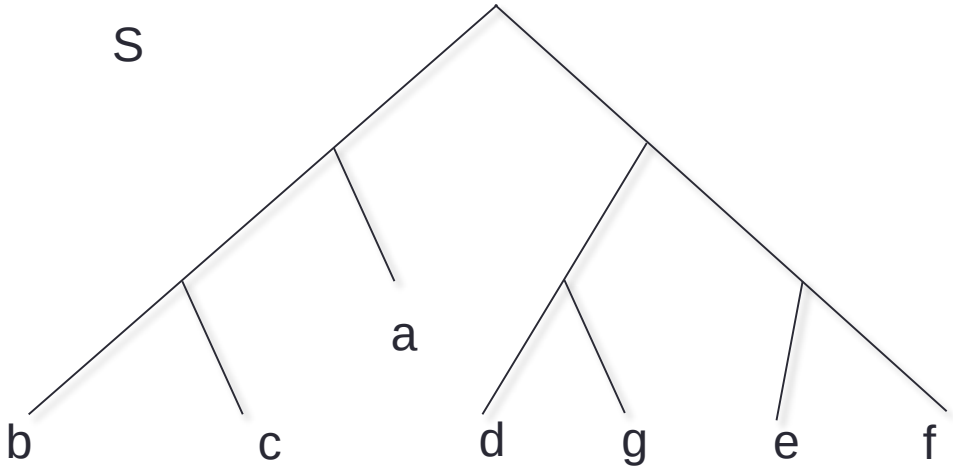
Consistency with a species tree S

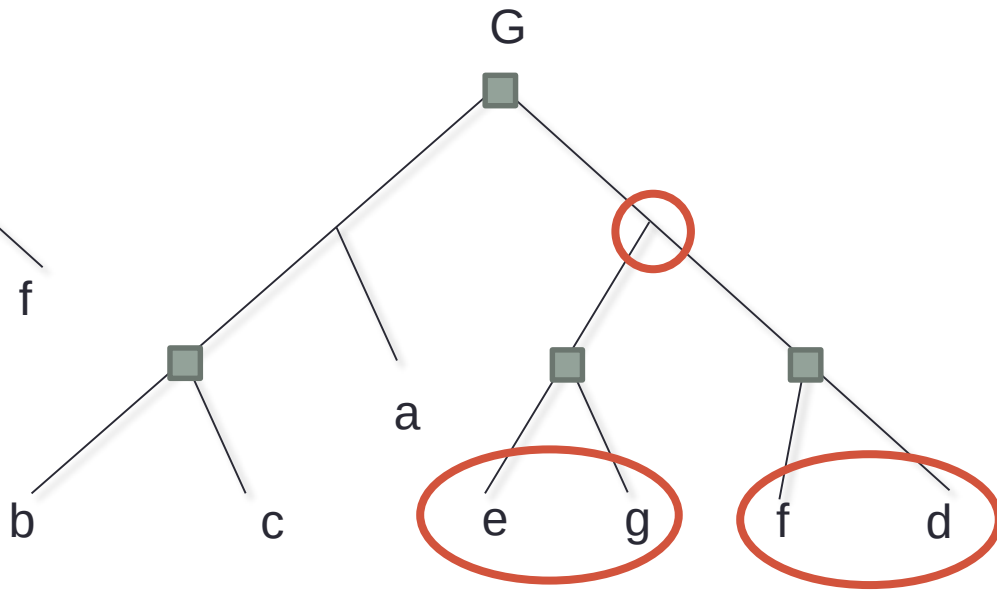




Consistency with a species tree

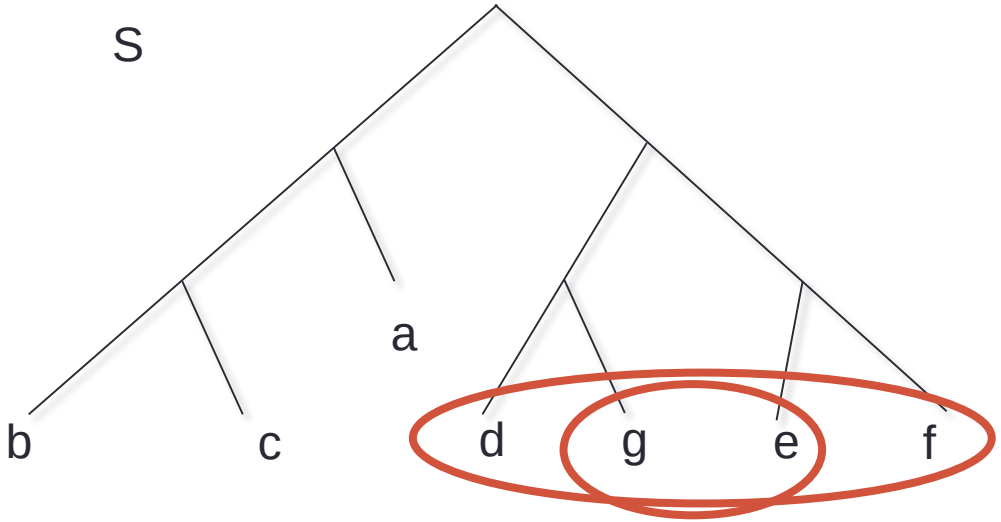
Consistency with S: If genes from species sets X,Y are separated by speciation in G, then species X, Y are separated in S.



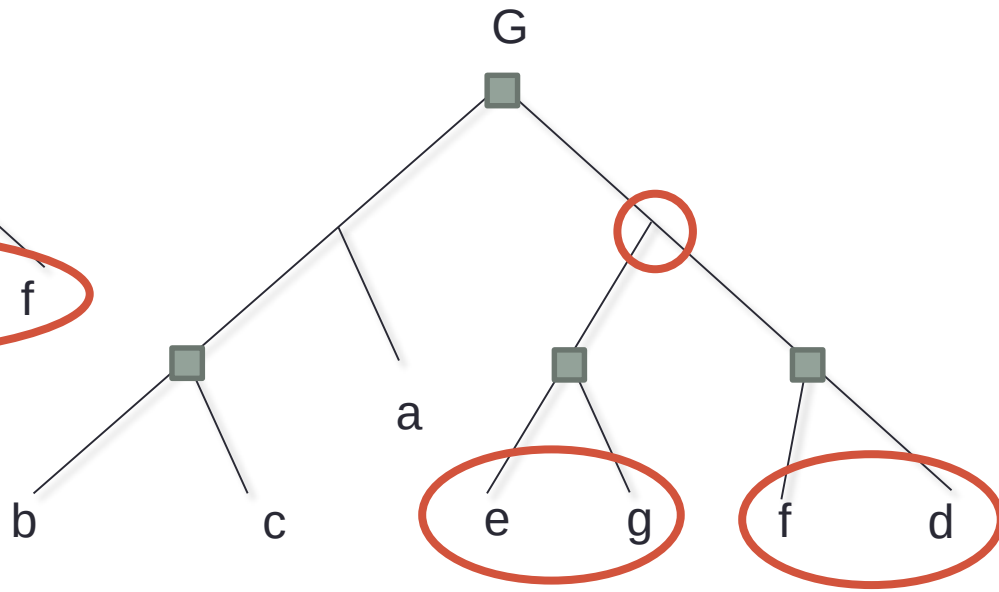


Consistency with a species tree

Consistency with S: If genes from species sets X,Y are separated by speciation in G, then species X, Y are separated in S.

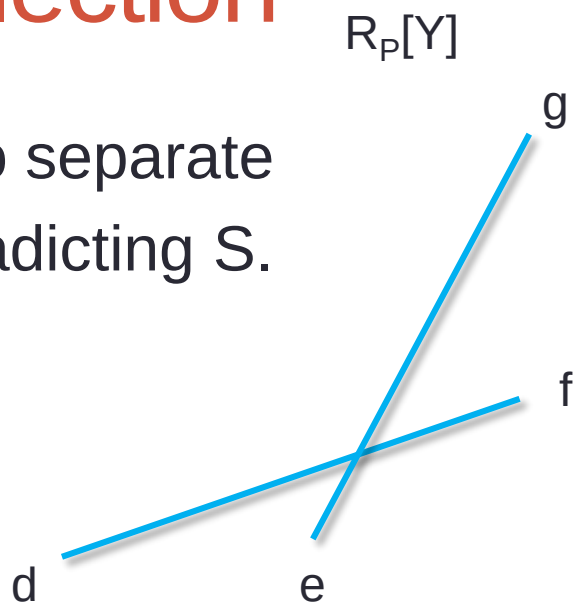


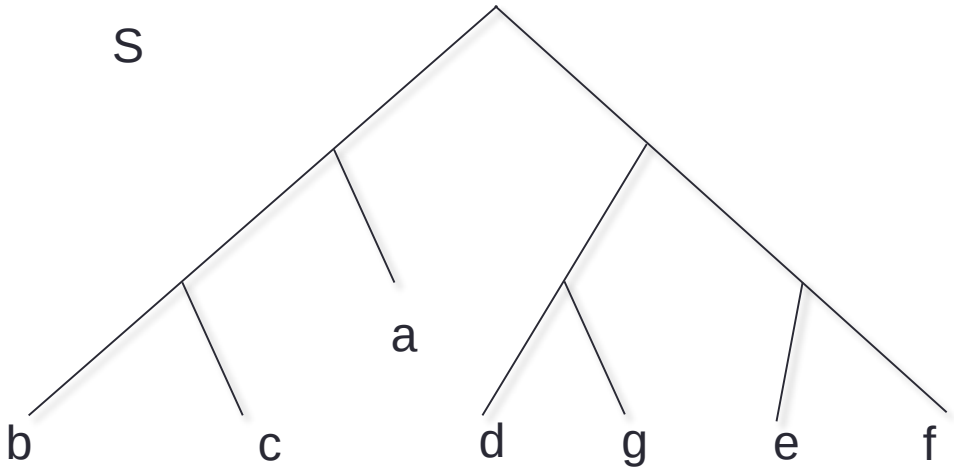
Inconsistent !

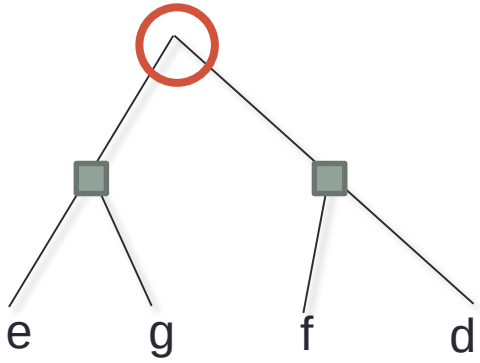


Careful component selection

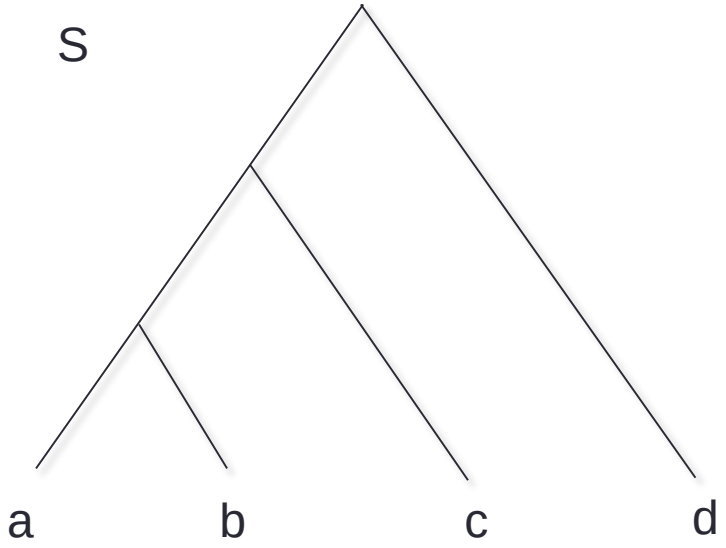
Problem: at this step Y, we chose to separate {e,g} from {f,d} by speciation, contradicting S.

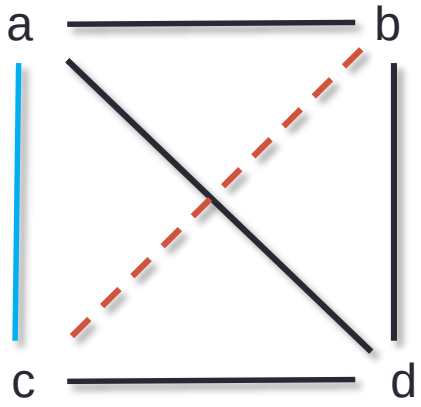




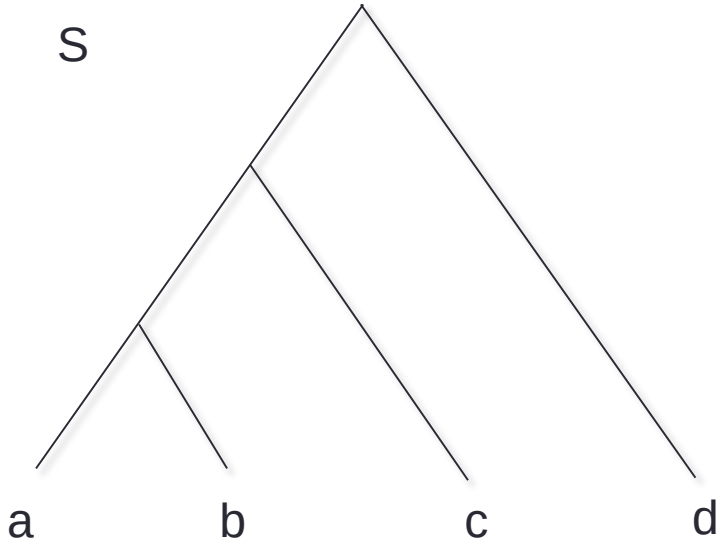


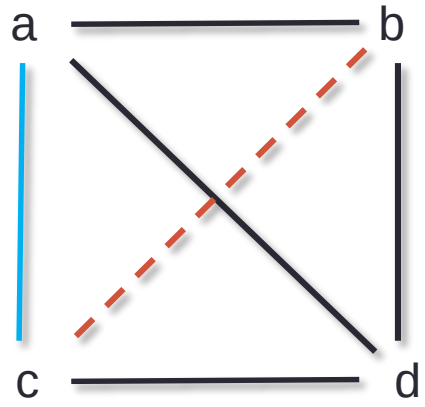
Careful component selection

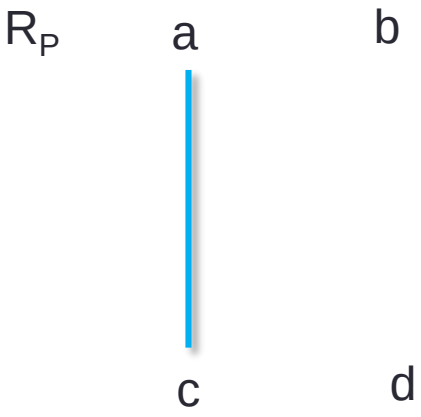




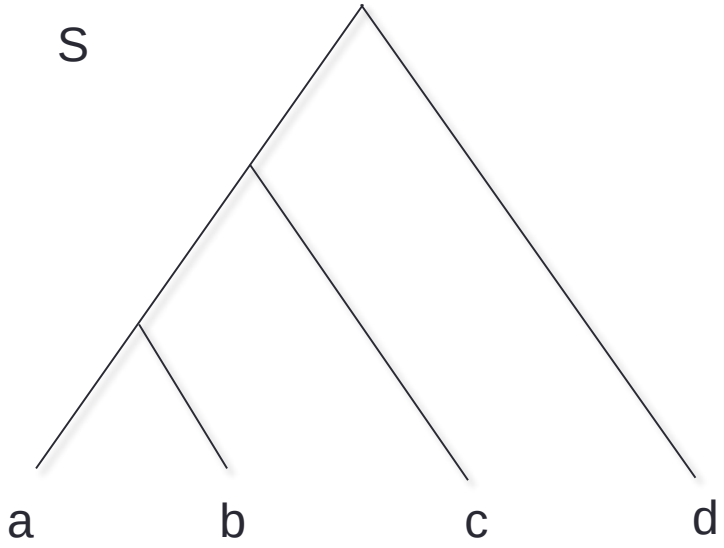
Careful component selection

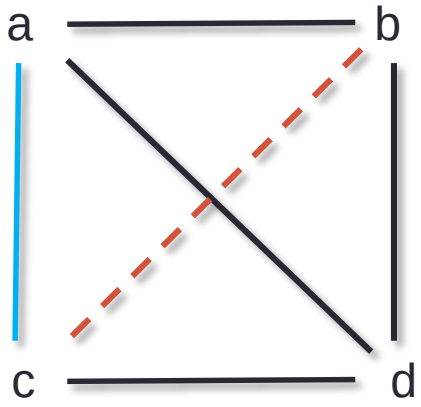


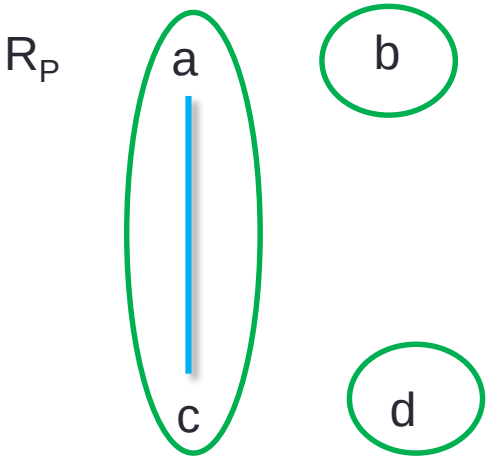




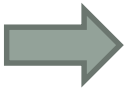
Careful component selection

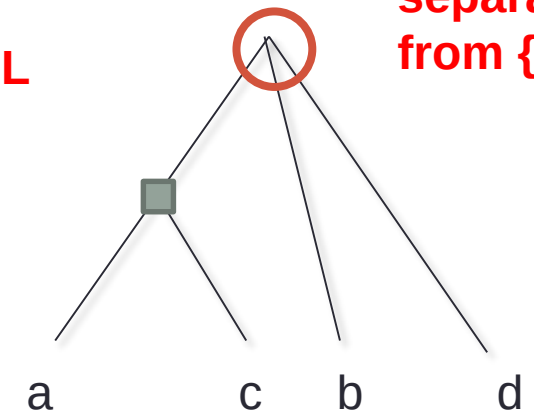






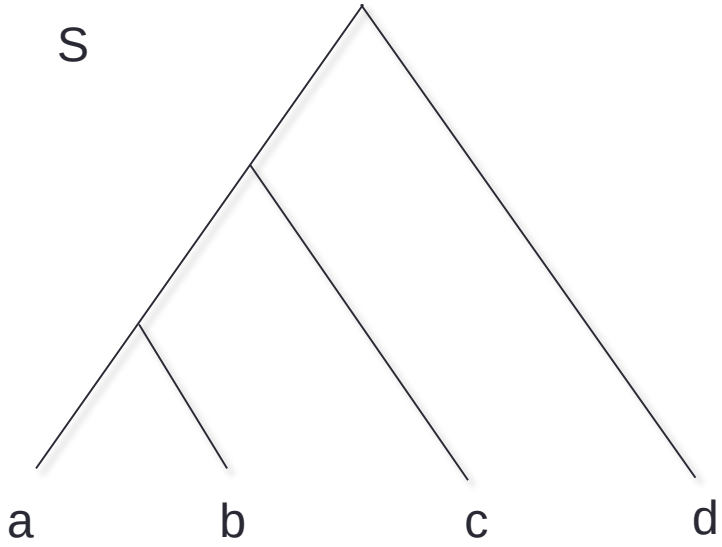
NOT CAREFUL

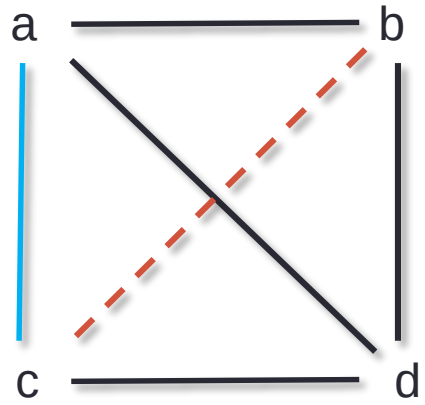


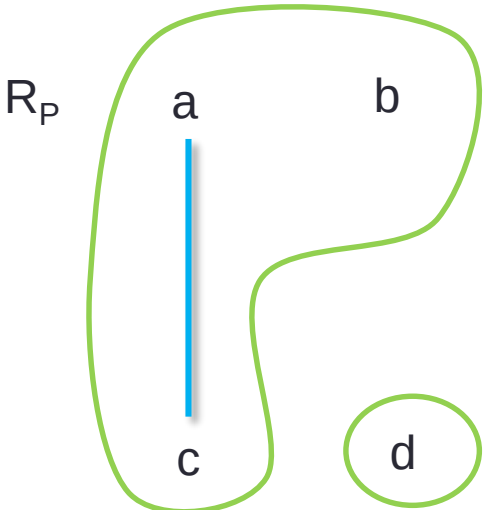


S does not separate {a,c} from {b}

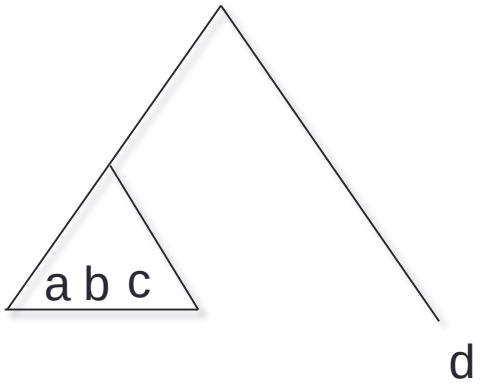
Careful component selection



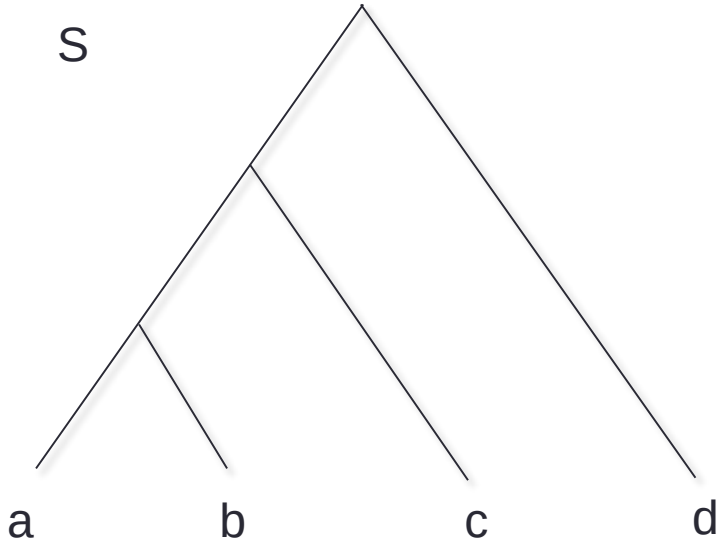


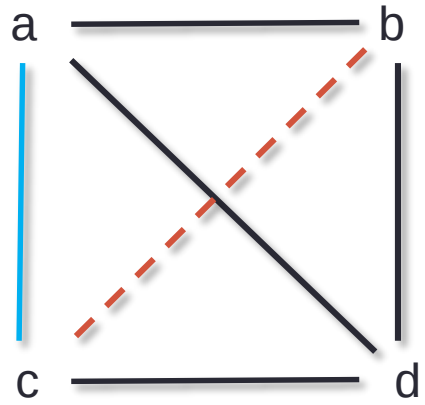


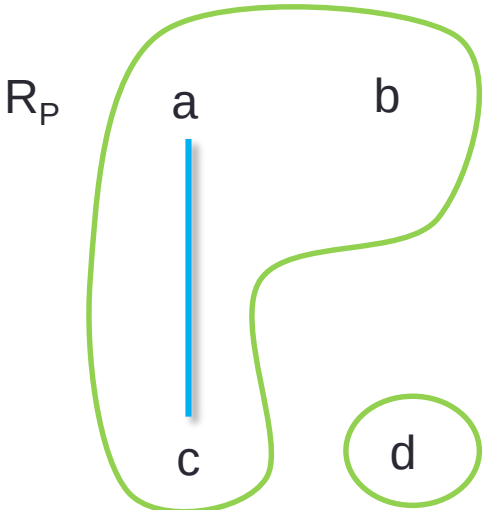
CAREFUL



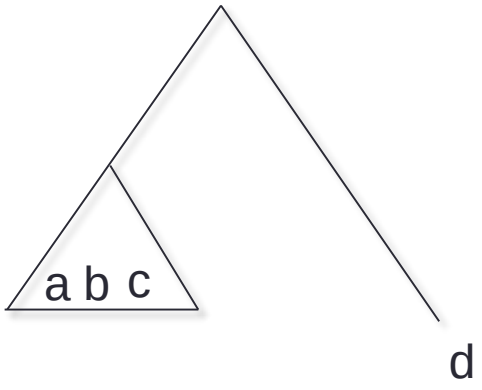
Careful component selection







CAREFUL



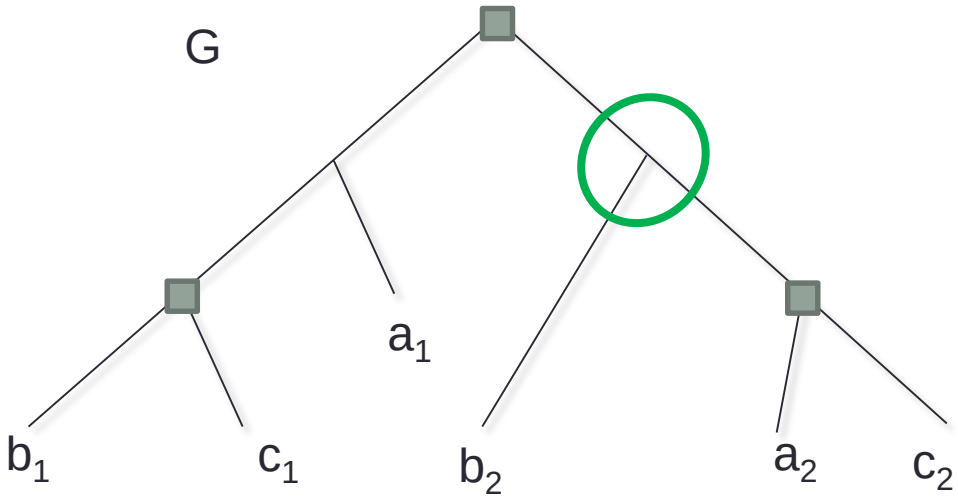
Consistency with S

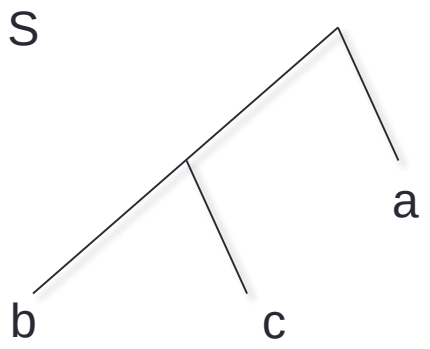
Theorem :

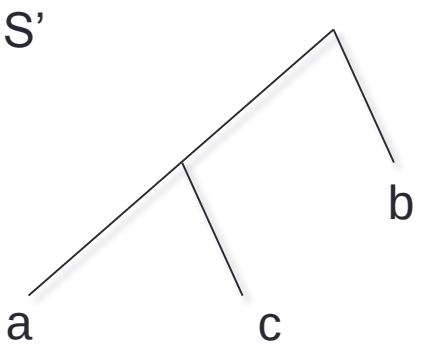
A relationship graph R is consistent with S iff **at least** one of the following holds :

1) R_O is disconnected, and each of its component is satisfiable
2) R_P is disconnected, its components admit a **non-trivial speciation partition P**, and each member of P is consistent with S

Self-consistency







Self-consistency

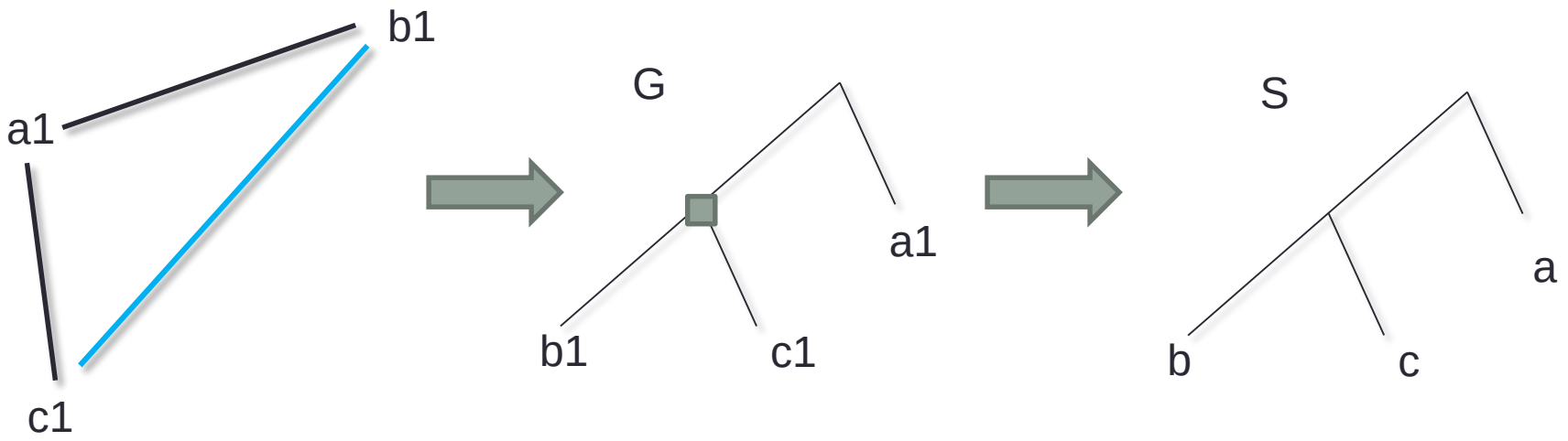
Is there some gene tree G that satsfies R,

such that **some** species tree S is consistent with G ?

The complexity of the problem is open…

Self-consistency

Suppose we have all relationships. Every triangle with exactly one blue edge forces a triplet in the gene tree, and consequently in the species tree.



Self-consistency

Theorem : a **full (no unknowns), satisfiable** relationship graph R is self-consistent (consistent with some species tree) iff

all triplets forced by one-blue-edge triangles can **all be displayed together** in the same species tree.

Self-consistency

Theorem : a **full (no unknowns), satisfiable** relationship graph R is self-consistent (consistent with some species tree) iff

all triplets forced by one-blue-edge triangles can **all be displayed together** in the same species tree.

Branch-and-bound algorithm with unknown edges :

Try both possibilities with every unknown edge e.

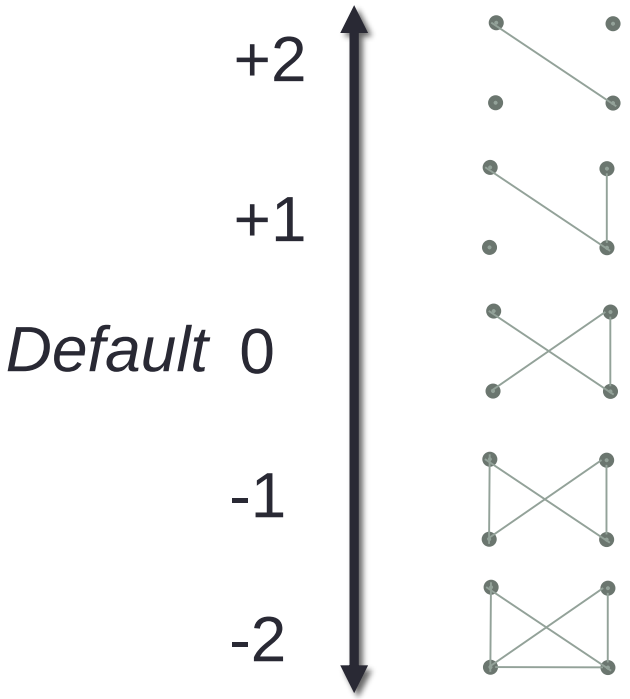
At every choice, run **BUILD** on the forced triplets.

If BUILD fails, don’t keep going and try some other choice.

Experiments

We looked at 265 inferred families from **ProteinOrtho**, under 5 parameter sets {-2, -1, 0, +1, +2}.

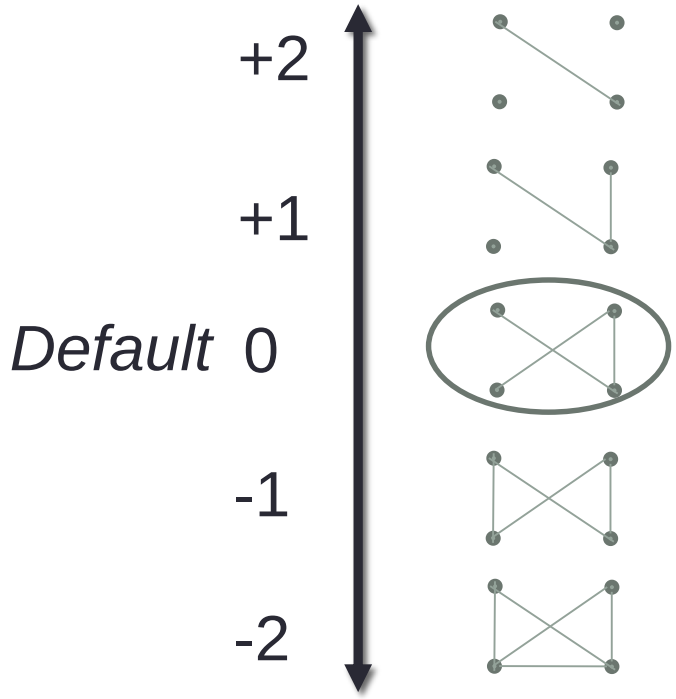
Stricter => Less orthologies



Looser => More orthologies

Experiments

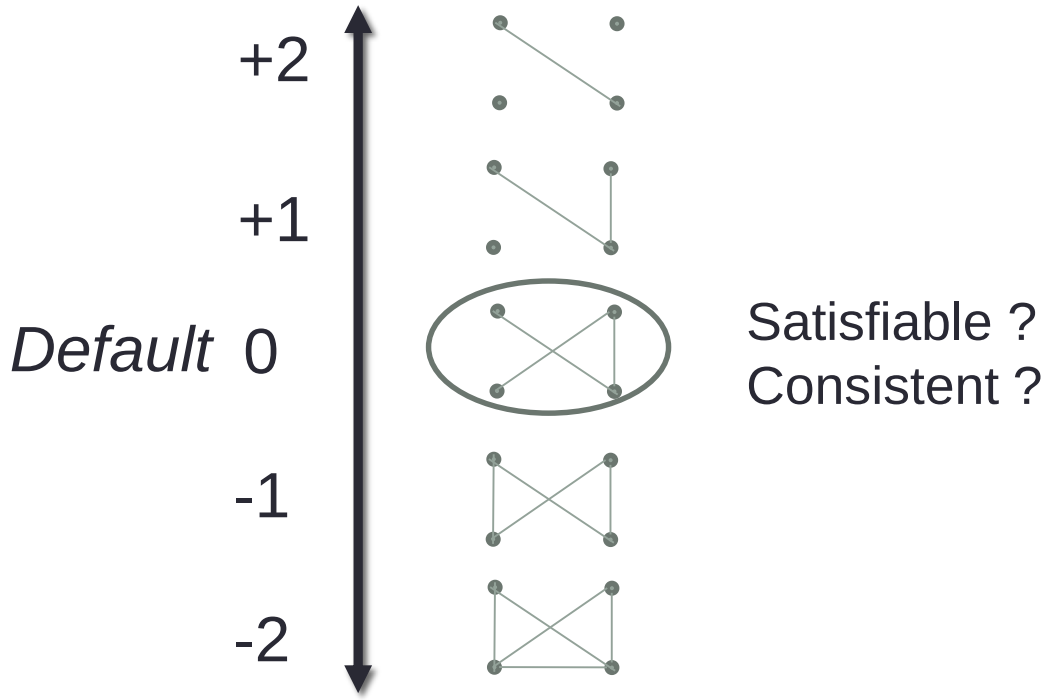
Stricter => Less orthologies



Looser => More orthologies

Experiments

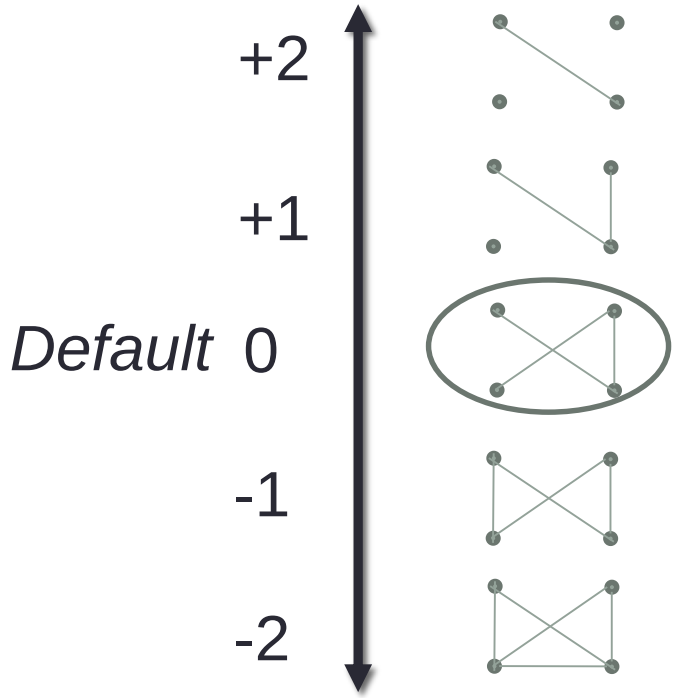
Stricter => Less orthologies



Looser => More orthologies

Experiments

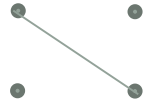
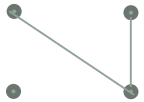
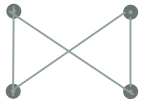
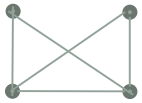
Stricter => Less orthologies



Satisfiable ? NO (~90% of families)
Consistent ? NO (~96% of families)

Looser => More orthologies

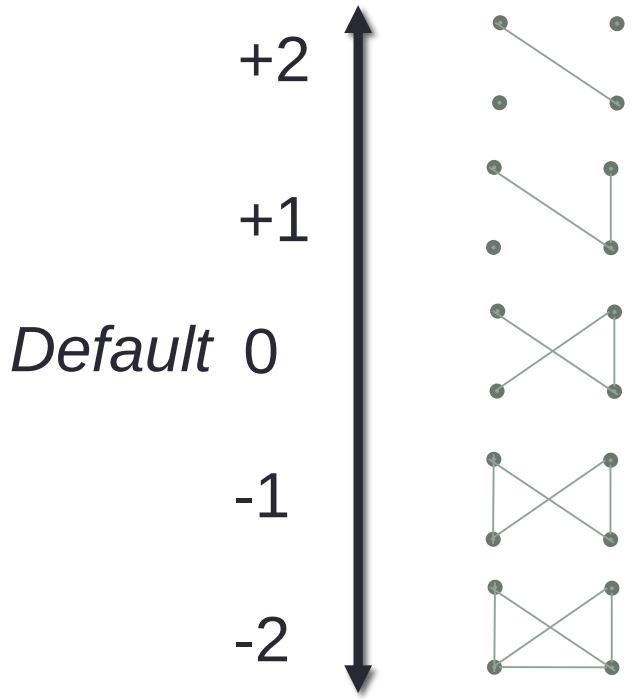
Experiments

Stricter => Less orthologies			NOT Satisfiable	NOT Consistent
	+2		80%	93%
	+1		82%	95%
Default	0		90%	96%
	-1		83%	95%
	-2		70%	89%

Looser => More orthologies

Experiments

Stricter => Less orthologies

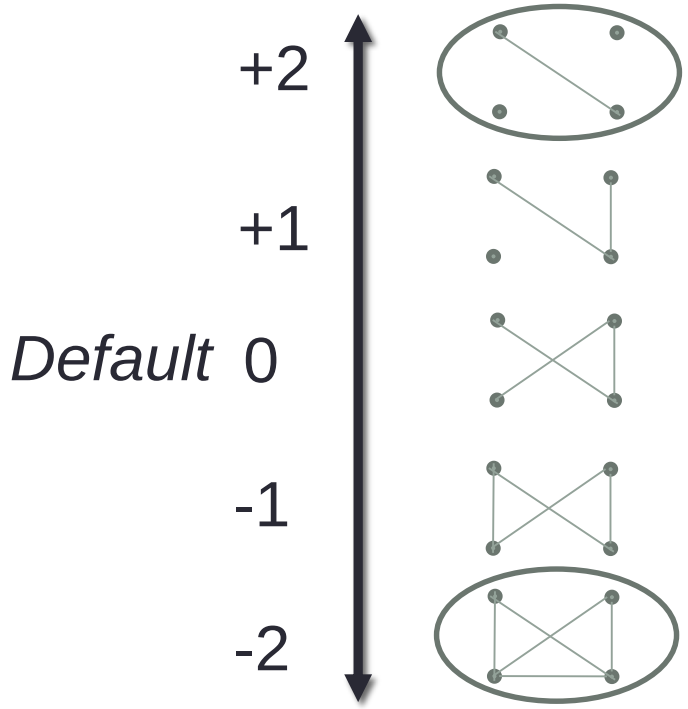


Looser => More orthologies

Can we get some robust relationships out of these ?

Experiments

Stricter => Less orthologies

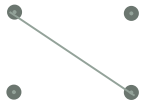


Looser => More orthologies

Can we get some robust relationships out of these ?

Experiments

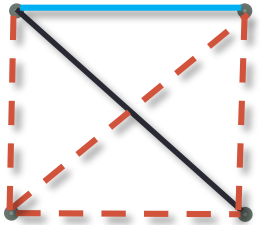
+2



-2

Keep the common orthologies and paralogies.

The rest is unknown.



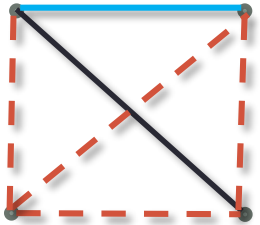
Experiments

+2

-2

Keep the common orthologies and paralogies.

The rest is unknown.



When combining +2/-2 as such, we find that these partial relationships are

satisfiable for 98% of families

consistent for 65% of families

On average, 42% of all possible relationships are known

Experiments

	NOT Satisfiable	NOT Consistent
-2/+2	1.9%	35.1%
-2/+1	2.6%	35.1%
-1/+1	4.2%	44.8%
-1/+2	4.1%	40.8%

Conclusion

- Gene tree correction
  - Given a set of consistent orthollogs/paralogs, modify G such that it exhibits the relationships
- Multiple solutions…how to choose one or list them ?
- Complexity $O(n^3)$ for satisfiability and consistency with a species tree
  - Can we do better ?
- Complexity of consistency : ????