Algorithms for the validation and correction of orthology relations

Manuel Lafond

University of Ottawa

Introduction

Gene trees, species trees

Duplication, speciation

Orthologs, paralogs, why?

Validation and correction of orthology relations

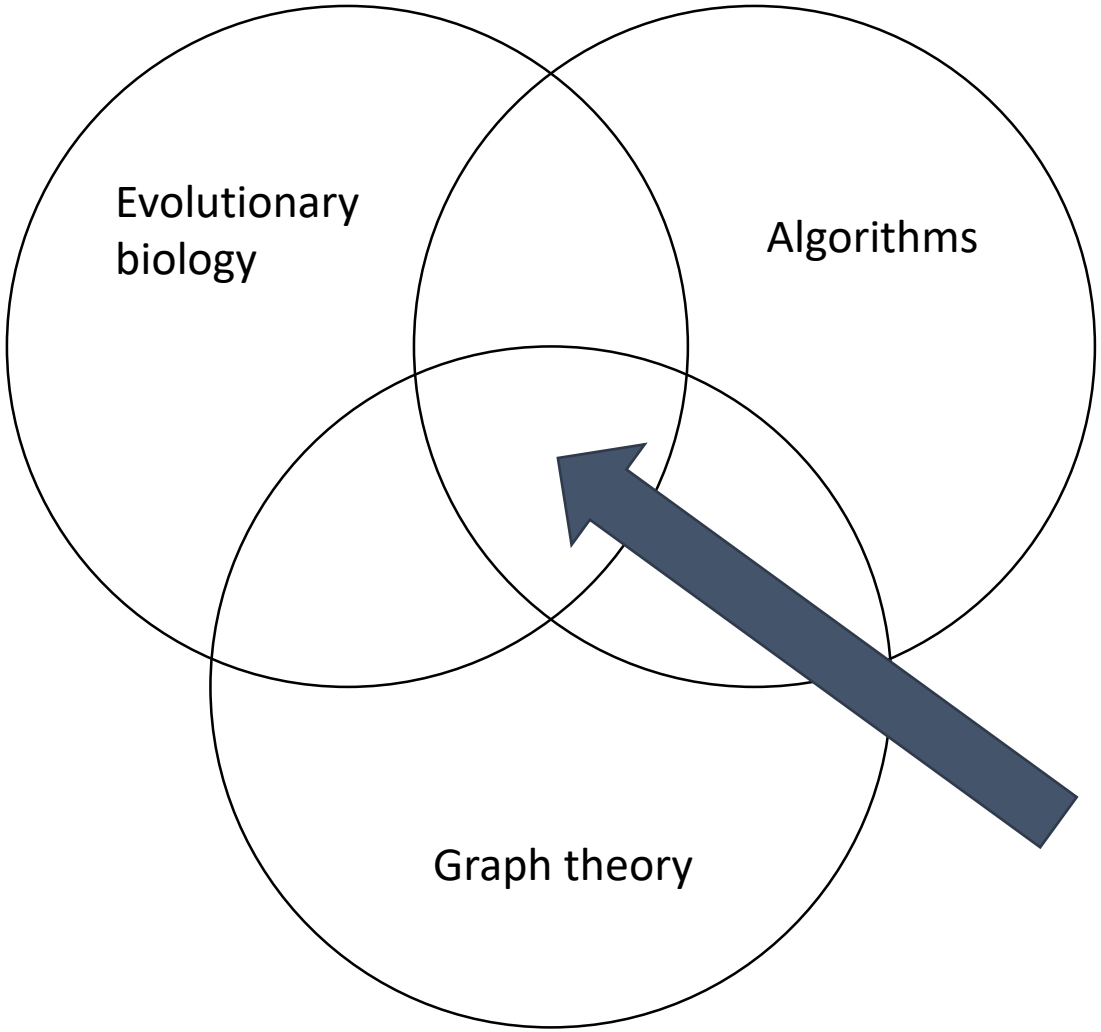
Cograph (P_4-free) characterization of valid relations

Modeling uncertain relations

Similarity graphs vs orthology graphs

Why they are not the same

How to deal with similarity graphs


Evolutionary biology
Algorithms
Graph theory

Introduction

Gene trees, species trees

Duplication, speciation

Orthologs, paralogs, why?

Validation and correction of orthology relations

Cograph (P_4-free) characterization of valid relations

Similarity graphs vs orthology graphs

Take some gene, say my favorite RPGR :

Retinitis pigmentosa GTPase regulator

Participates in eye coloring.

What is the **history** of RPGR ?

Almost all vertebrates have a copy of this gene. Some have more than one. Some don't have it.

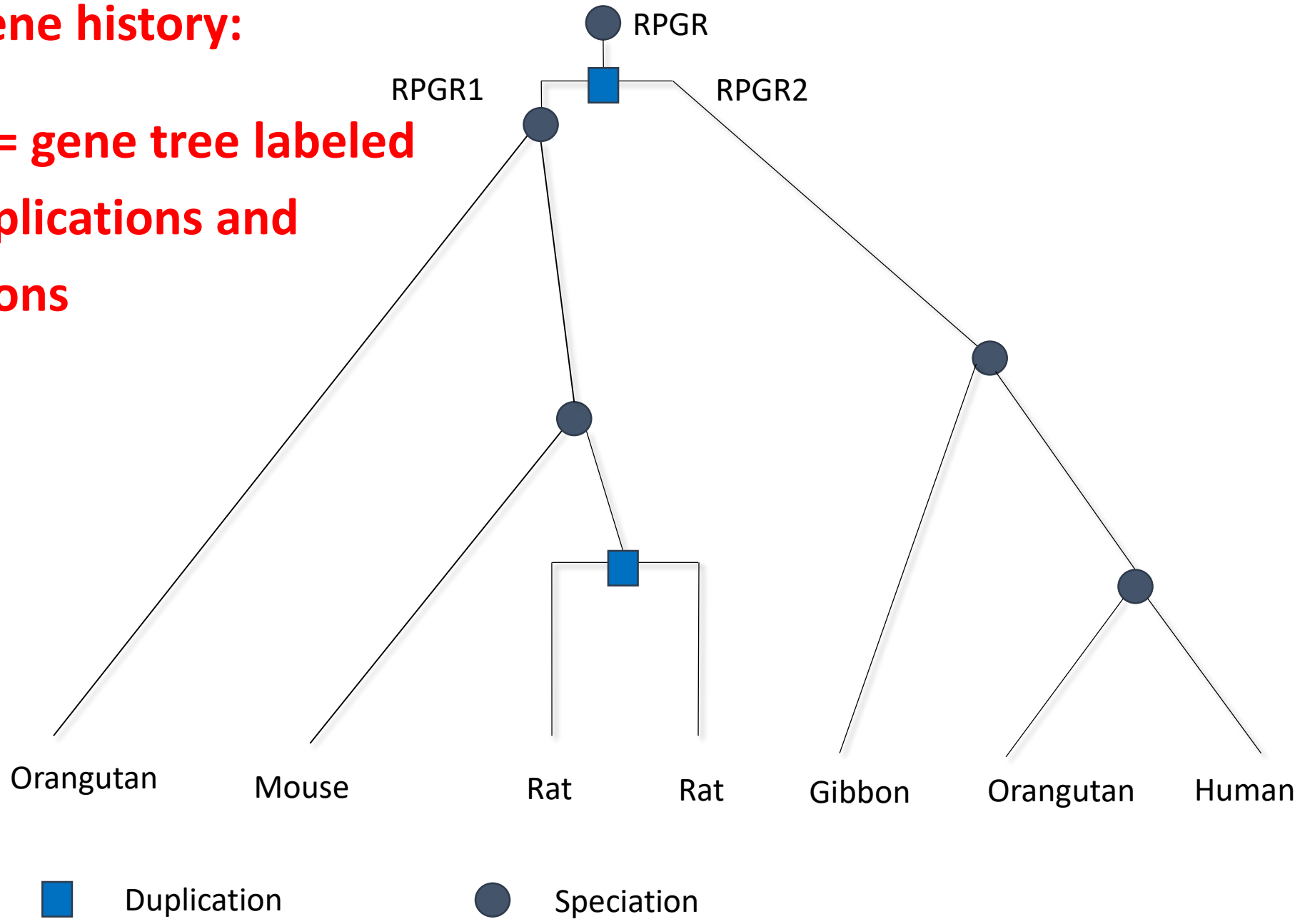
What happened exactly?

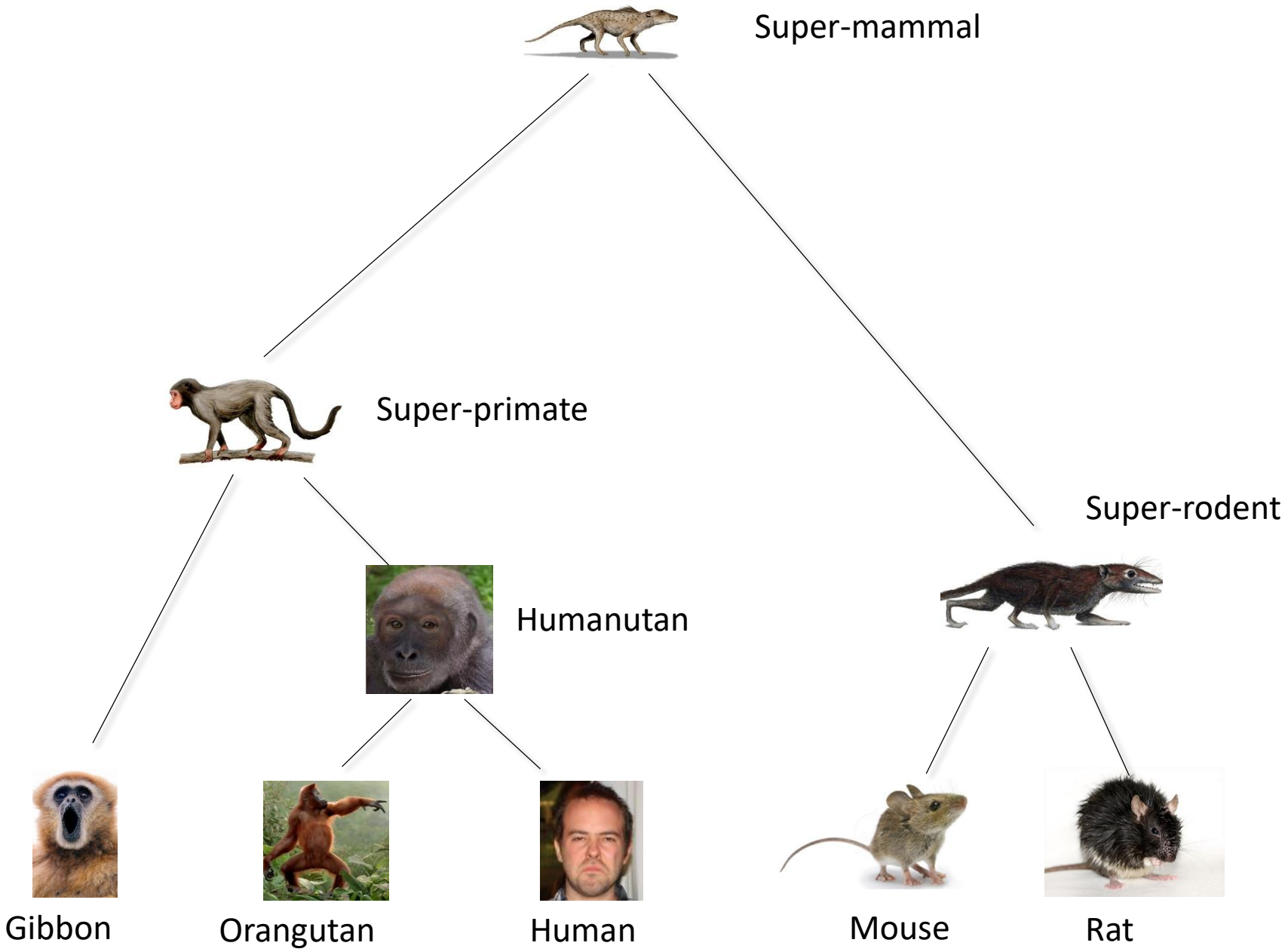
A gene can be :

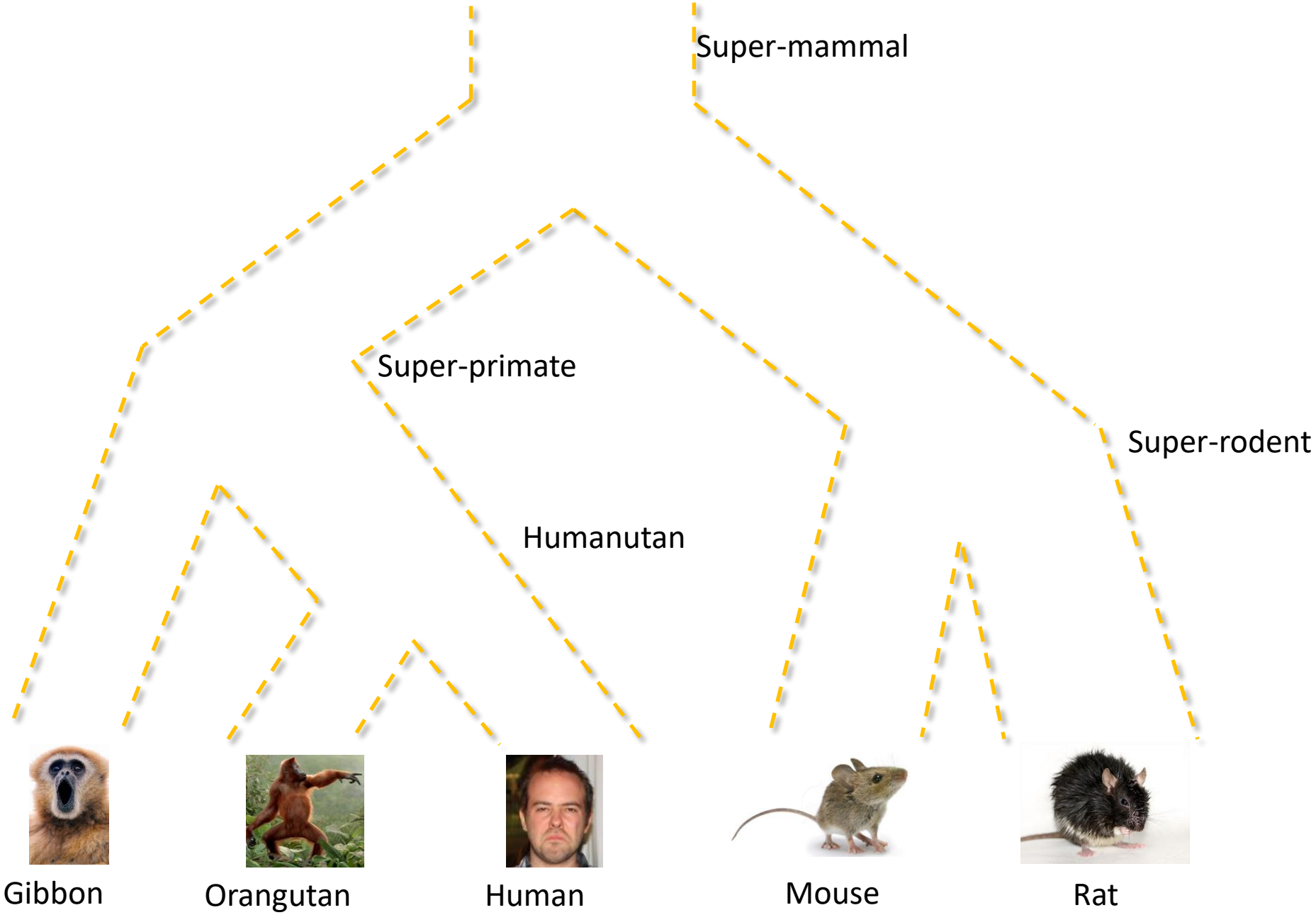
- **Transmitted** to descending species by **speciation**
- **Duplicated**
- **Lost**

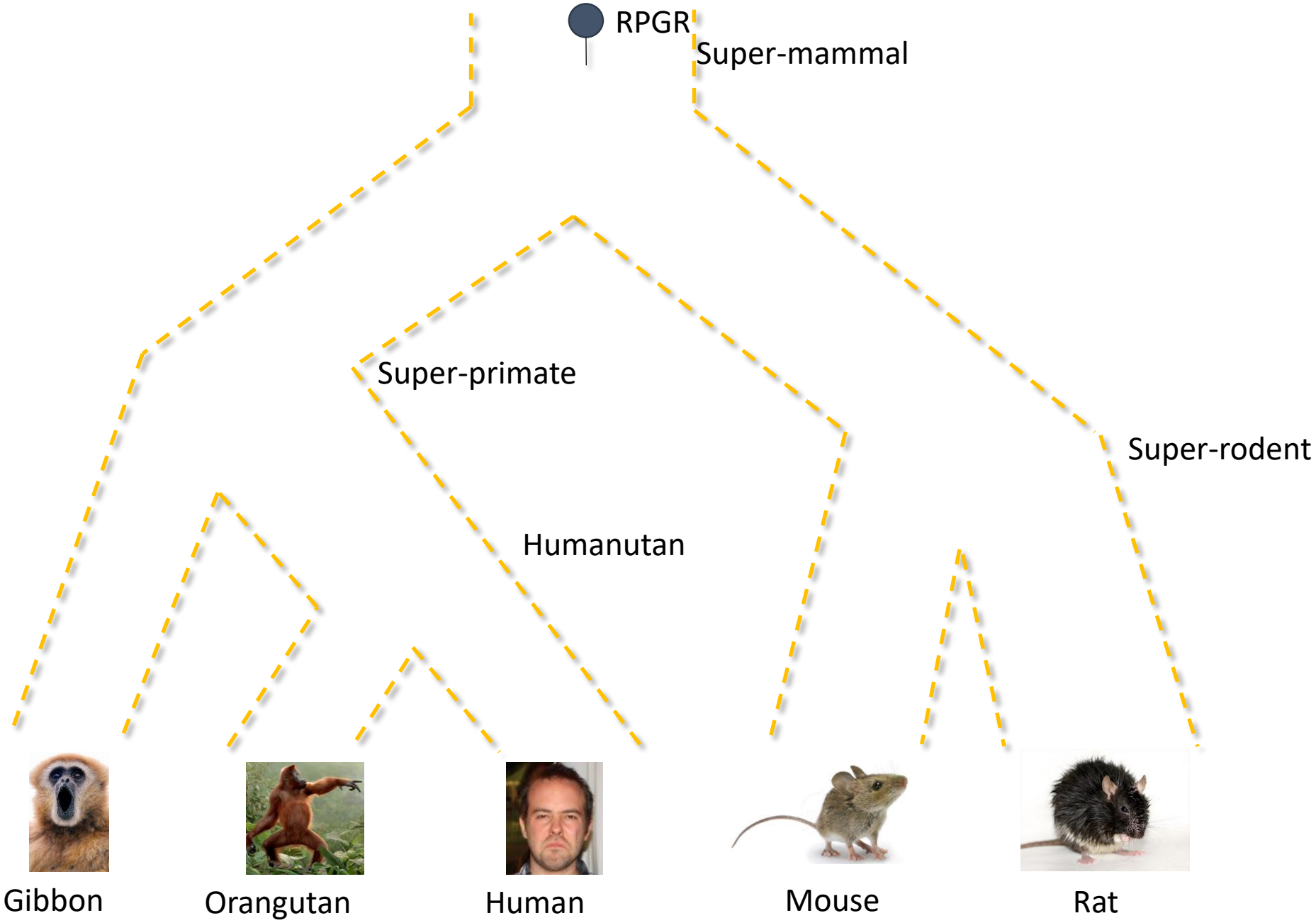
RPGR gene history:

History = gene tree labeled with duplications and speciations

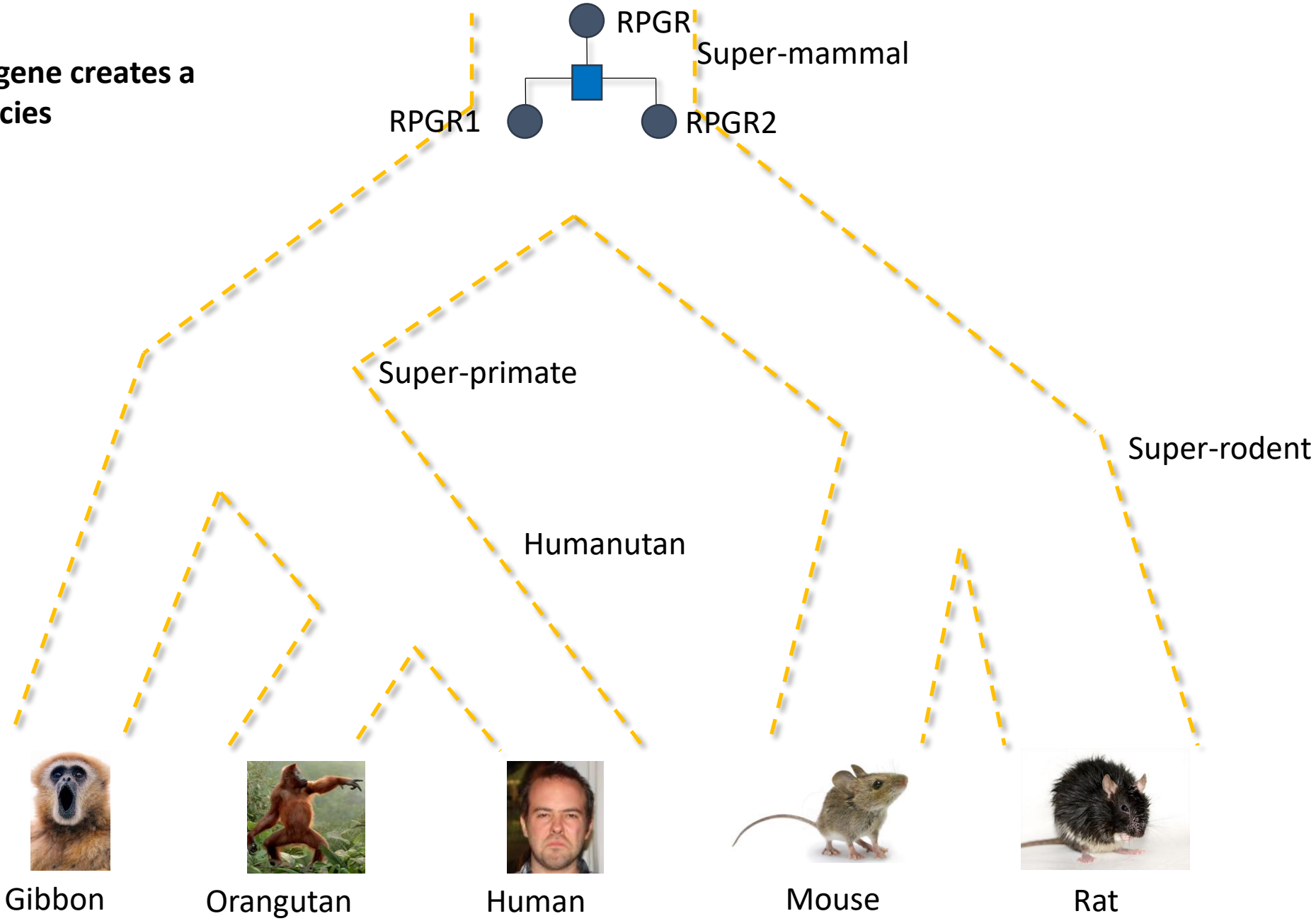



Super-mammal
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

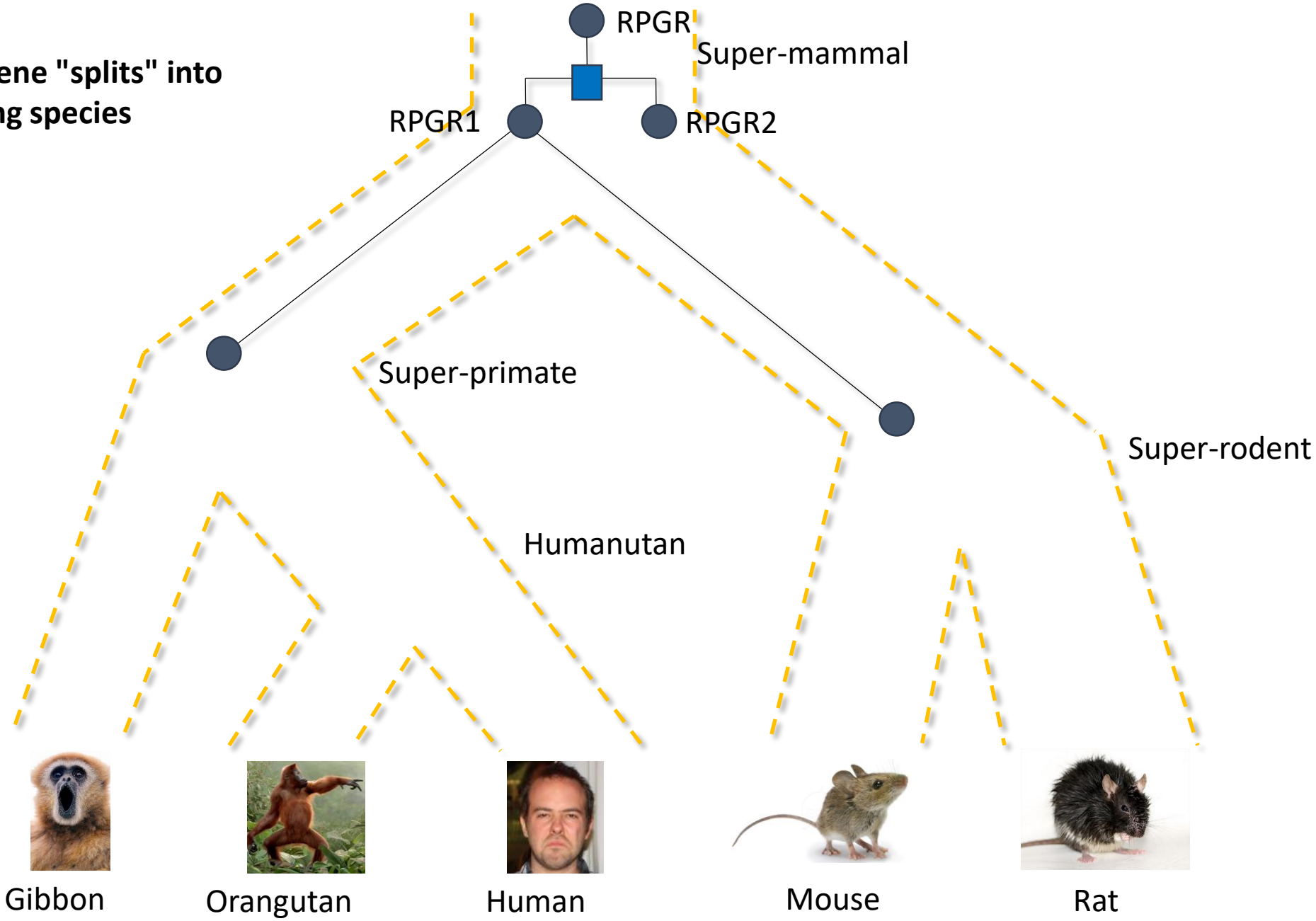

Super-mammal
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

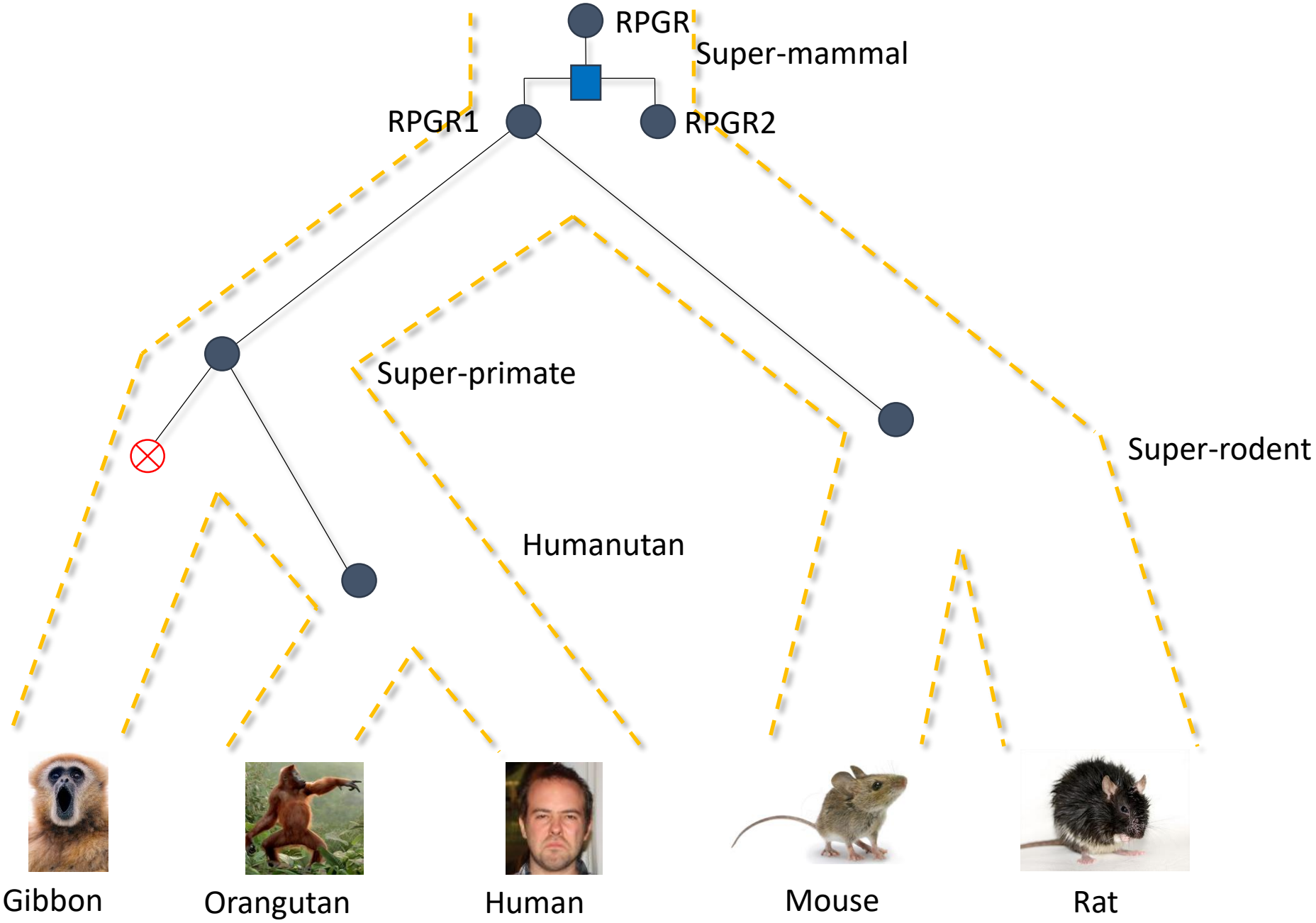

RPGR
Super-mammal
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

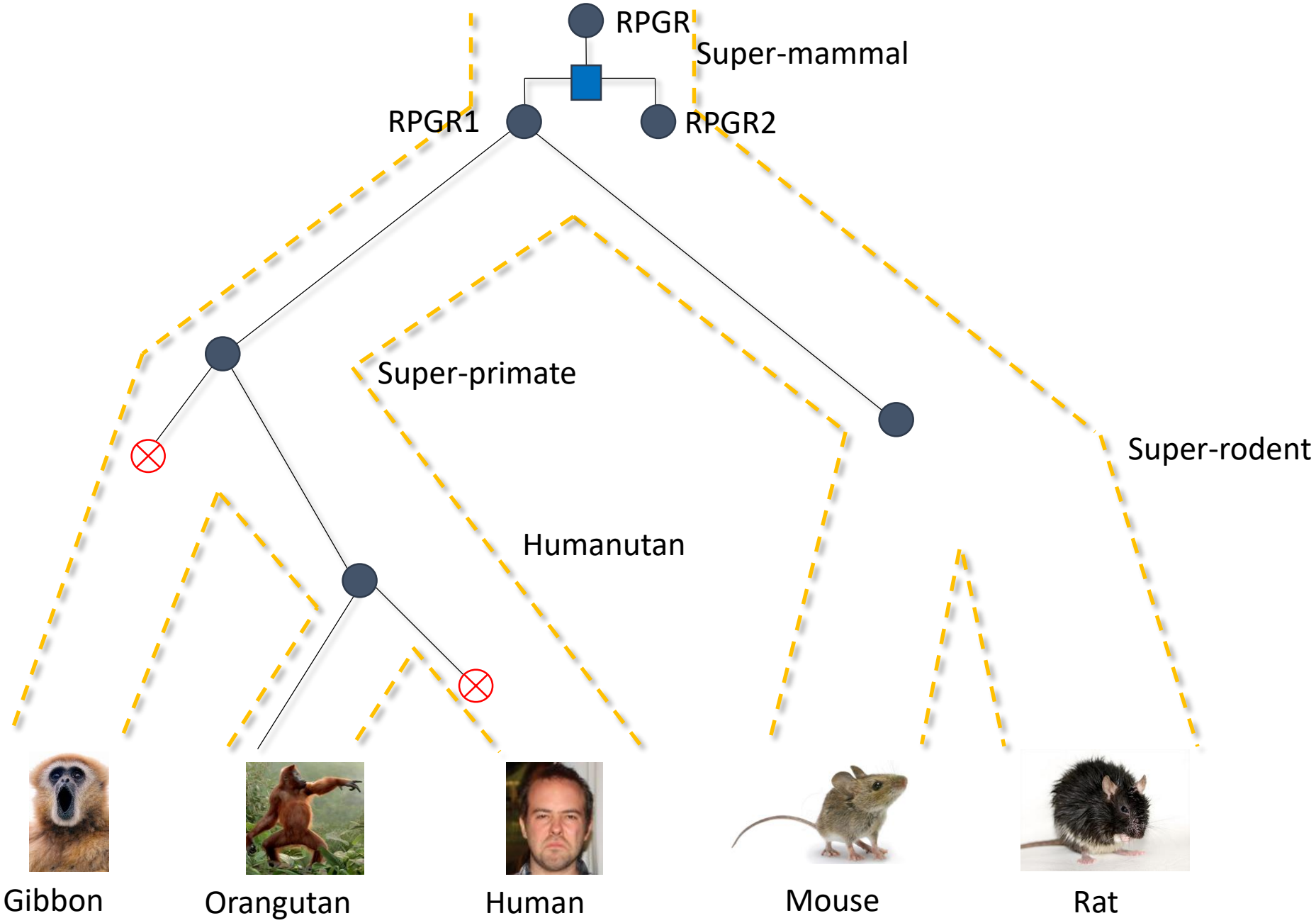
Duplication = gene creates a copy in its species

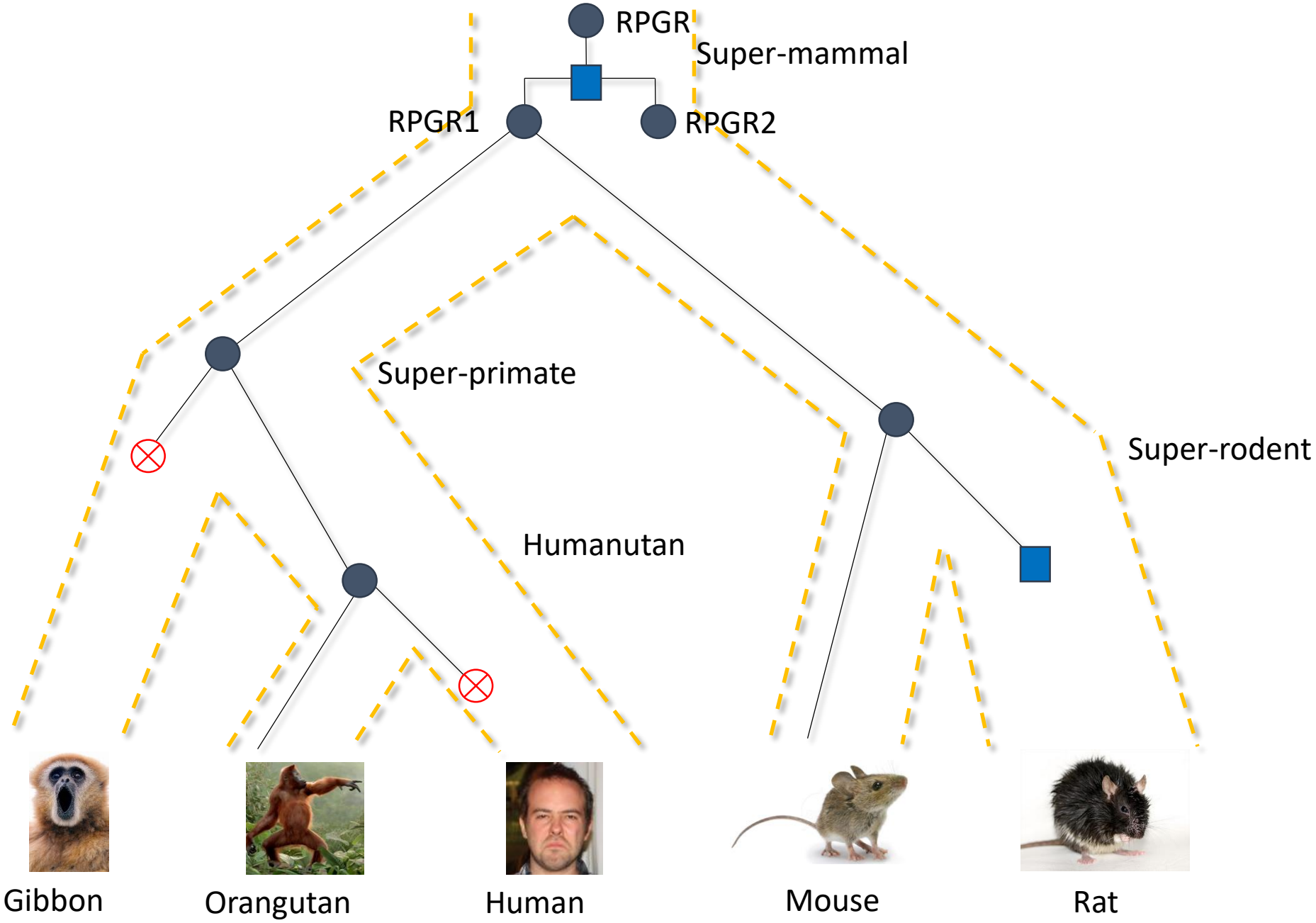


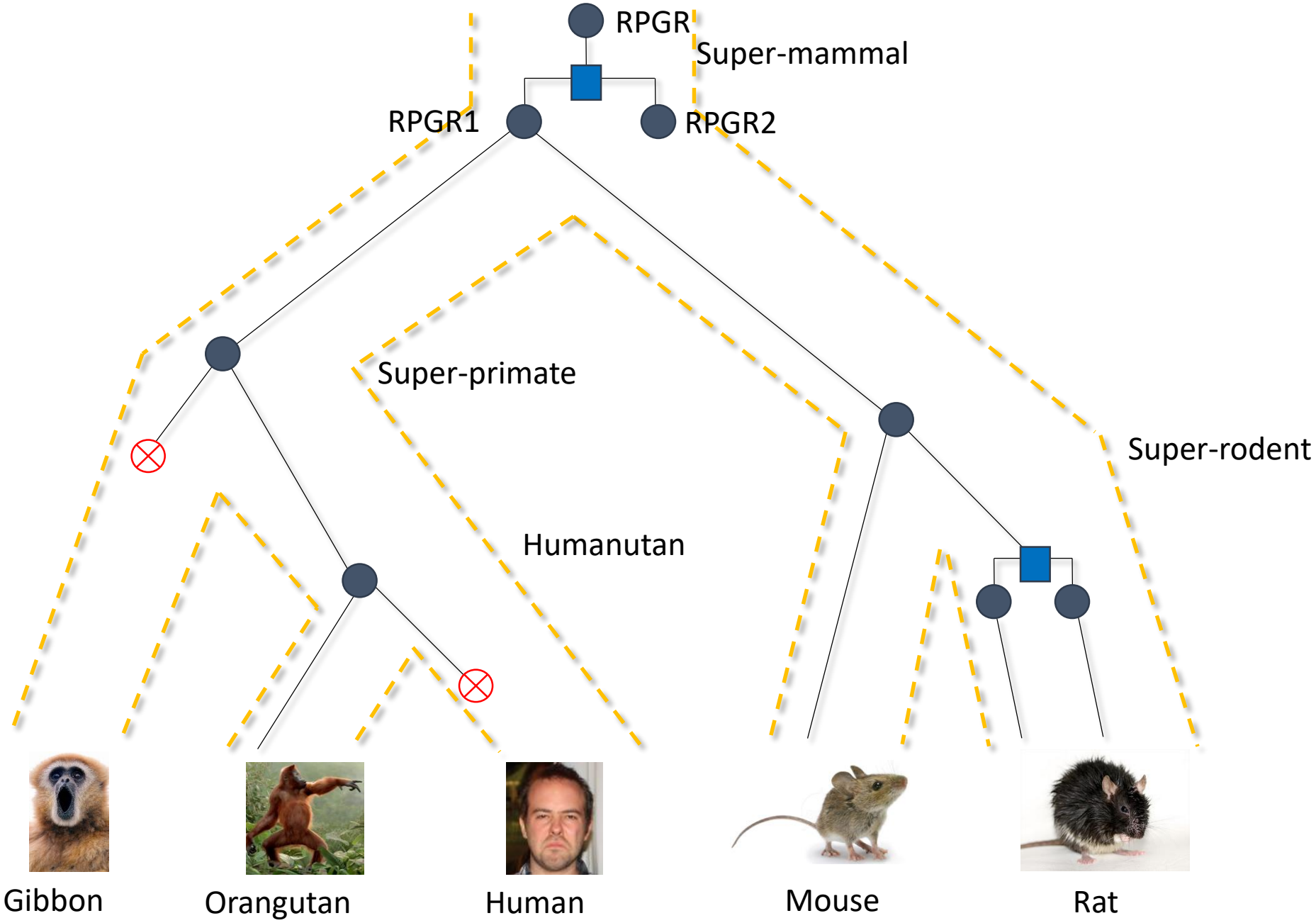
Speciation = gene "splits" into two descending species

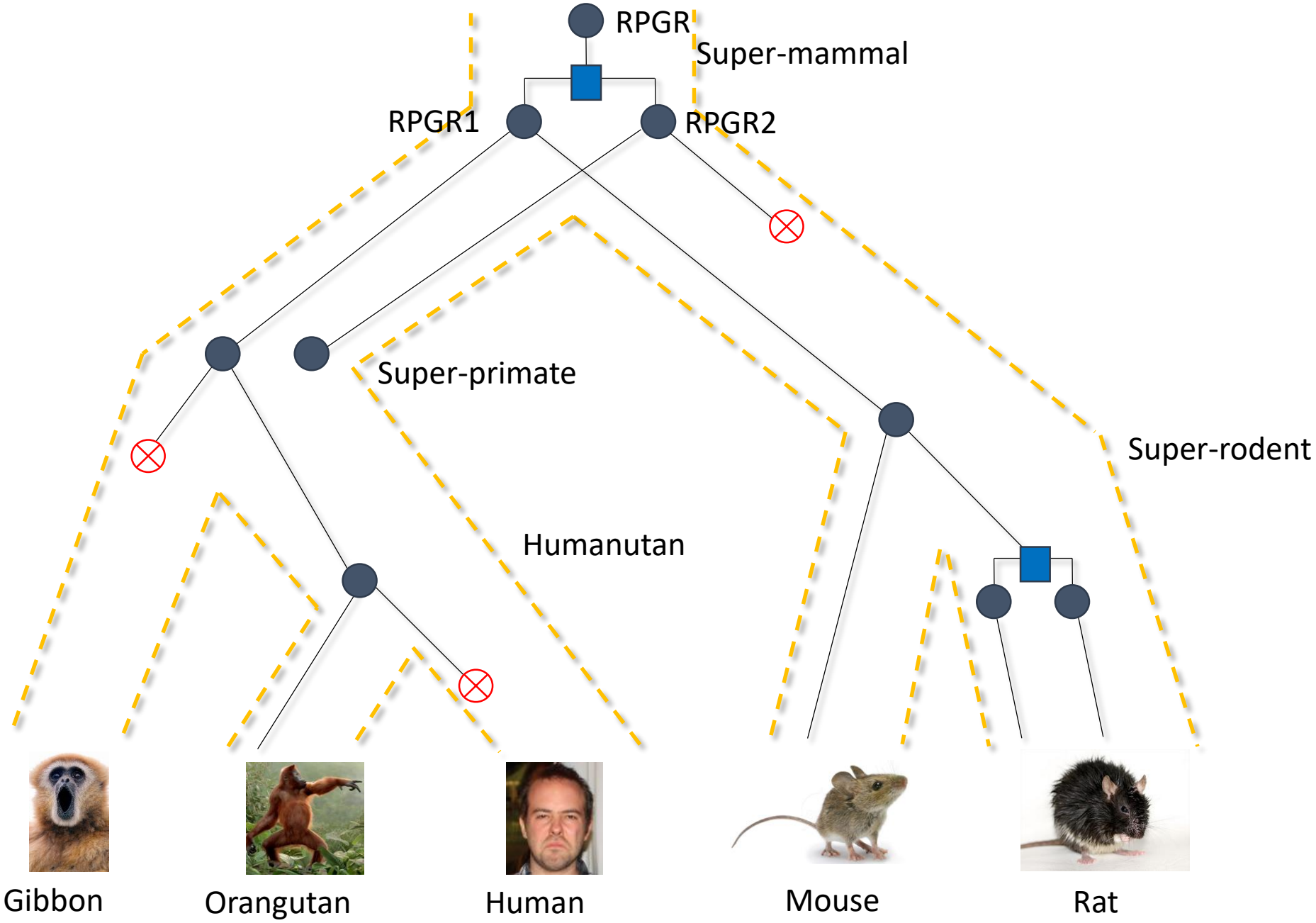


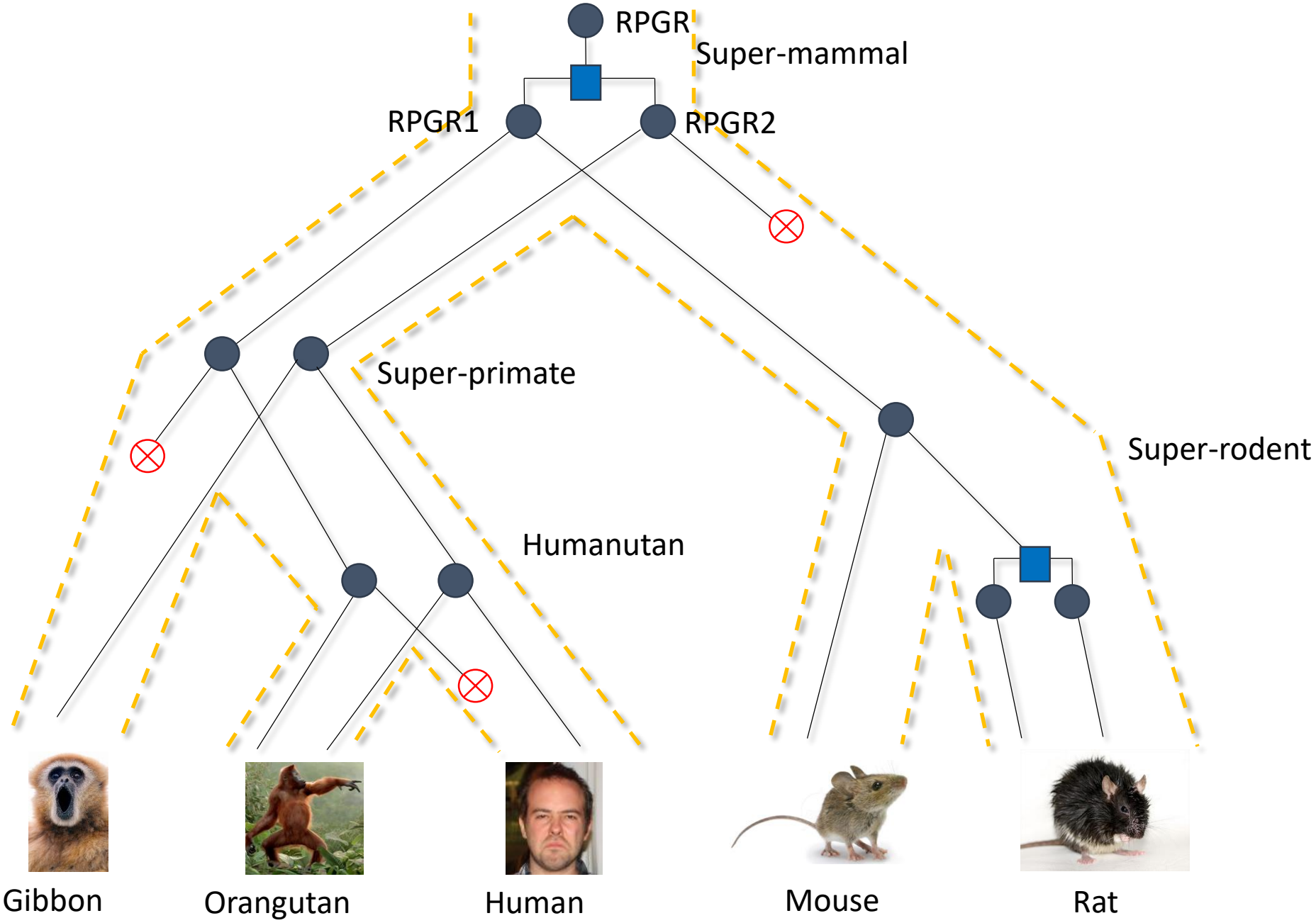

RPGR
Super-mammal
RPGR1
RPGR2
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat


RPGR
Super-mammal
RPGR1
RPGR2
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

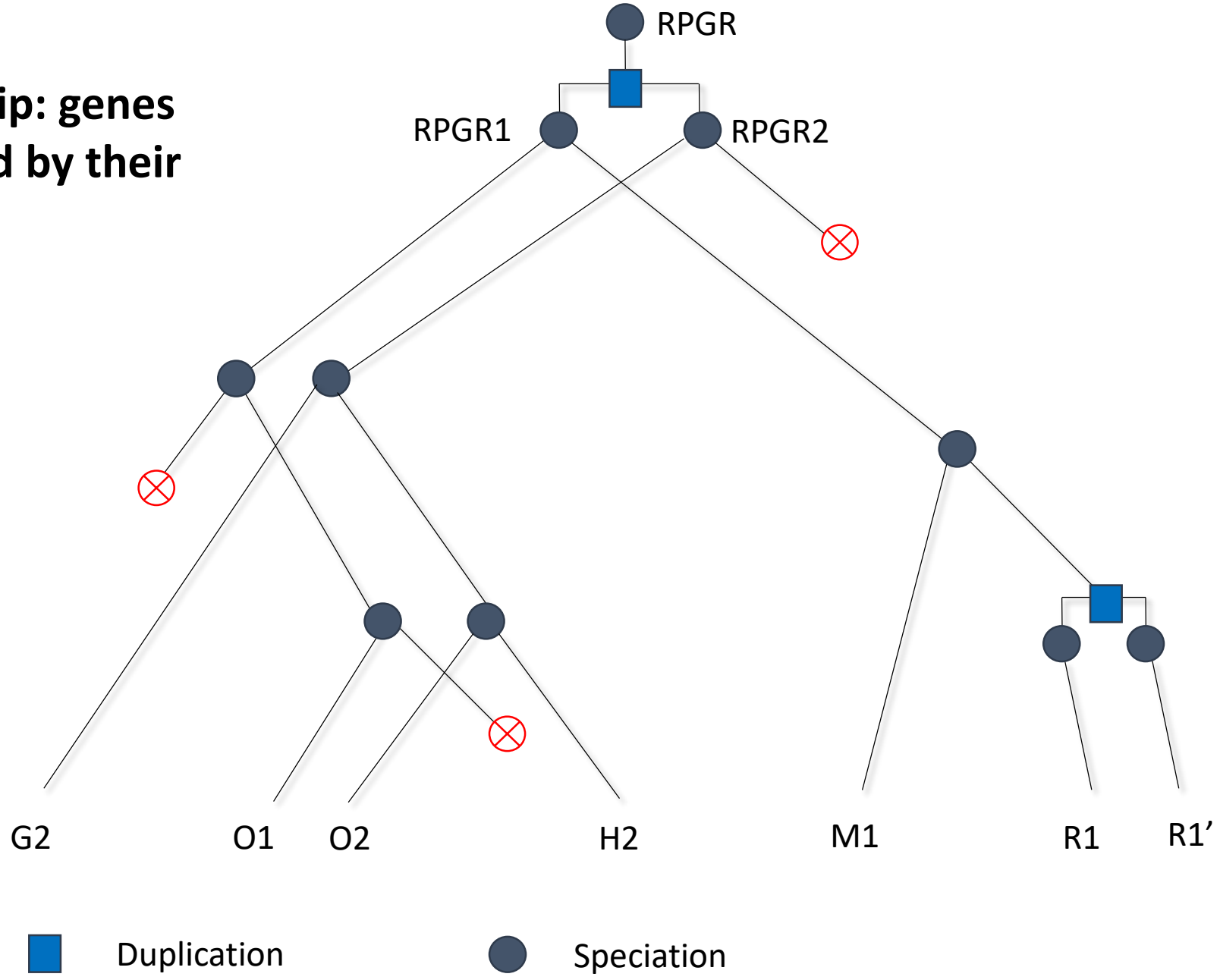

RPGR
Super-mammal
RPGR1
RPGR2
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

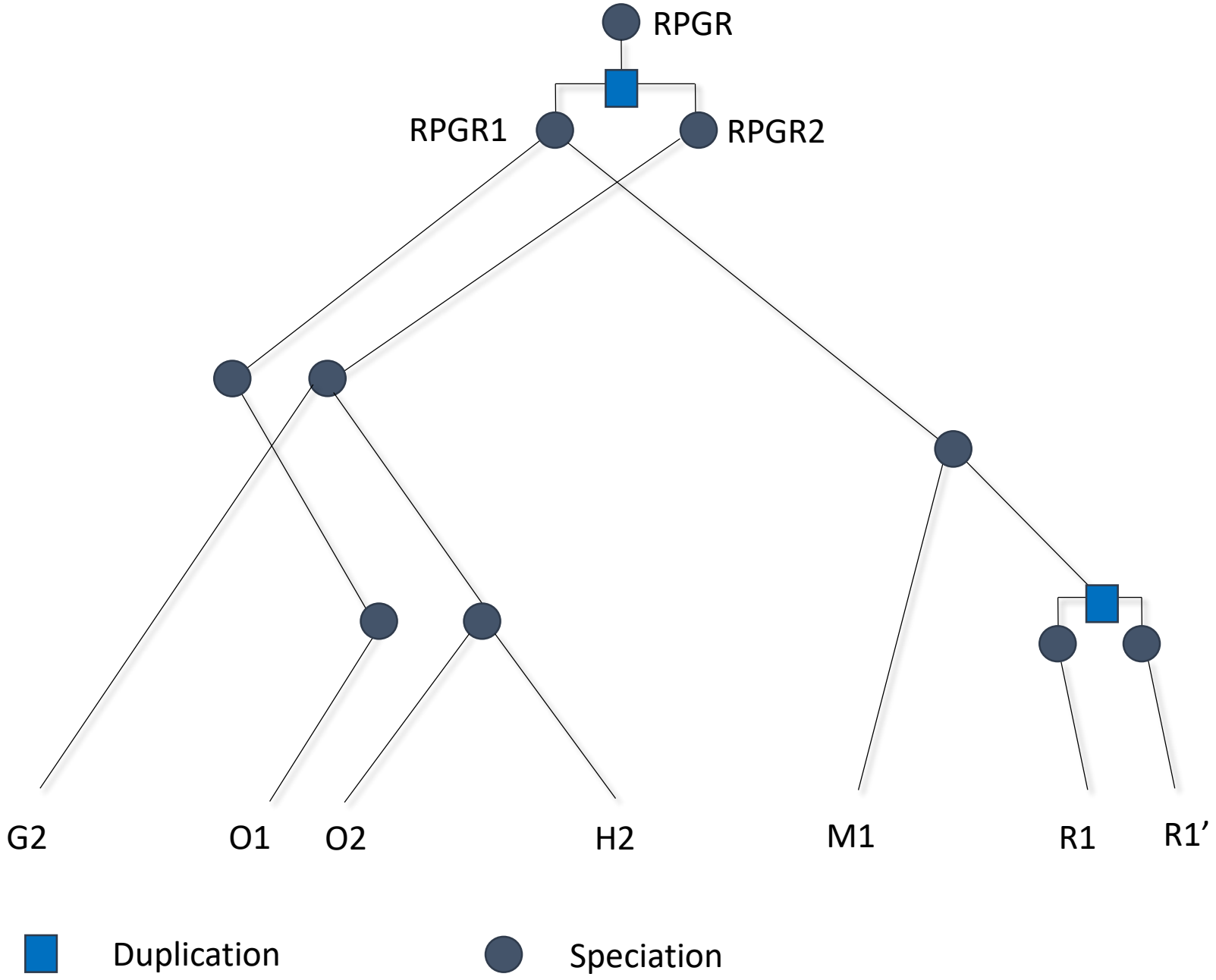

RPGR
Super-mammal
RPGR1
RPGR2
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

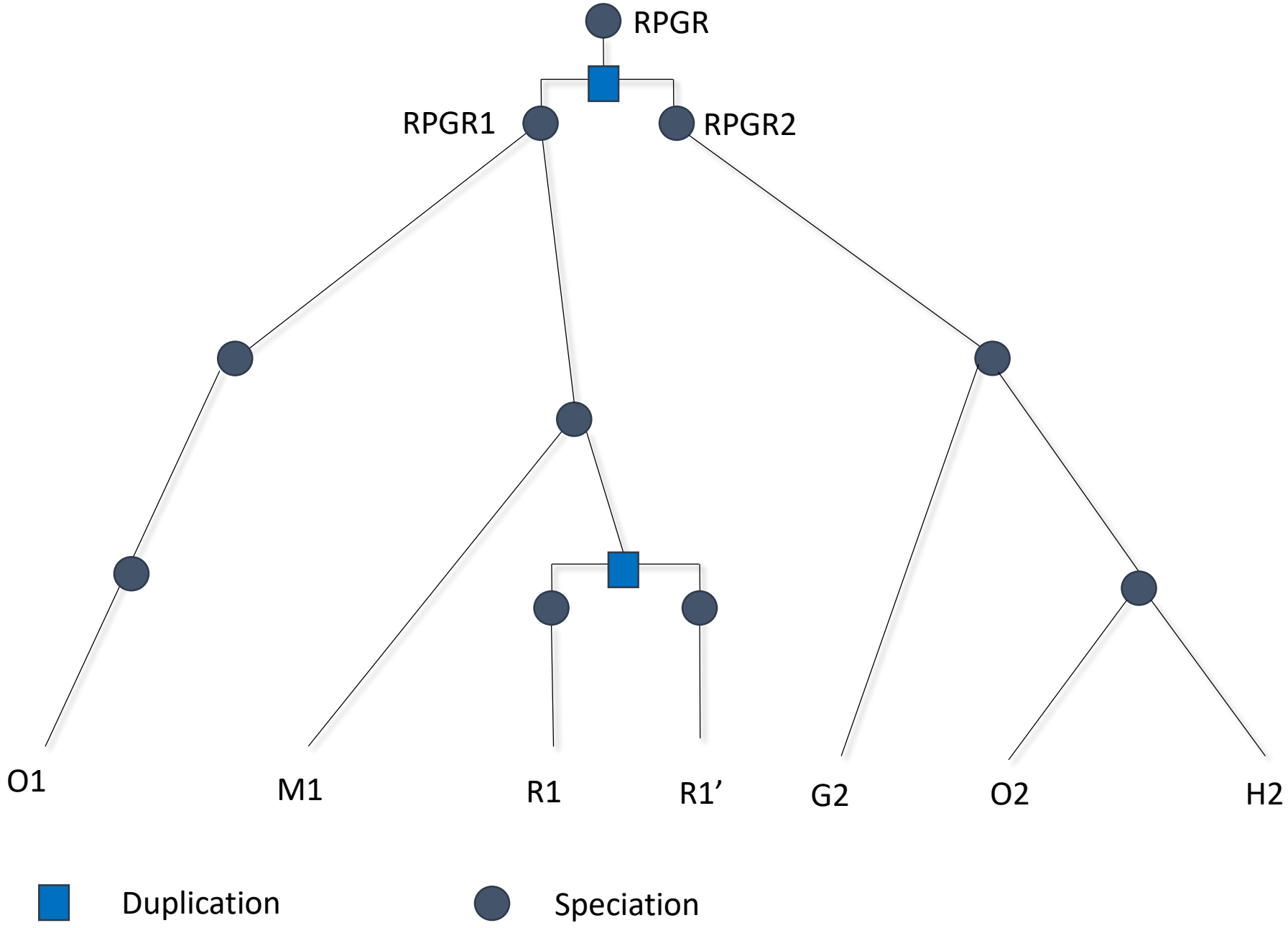

RPGR
Super-mammal
RPGR1
RPGR2
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

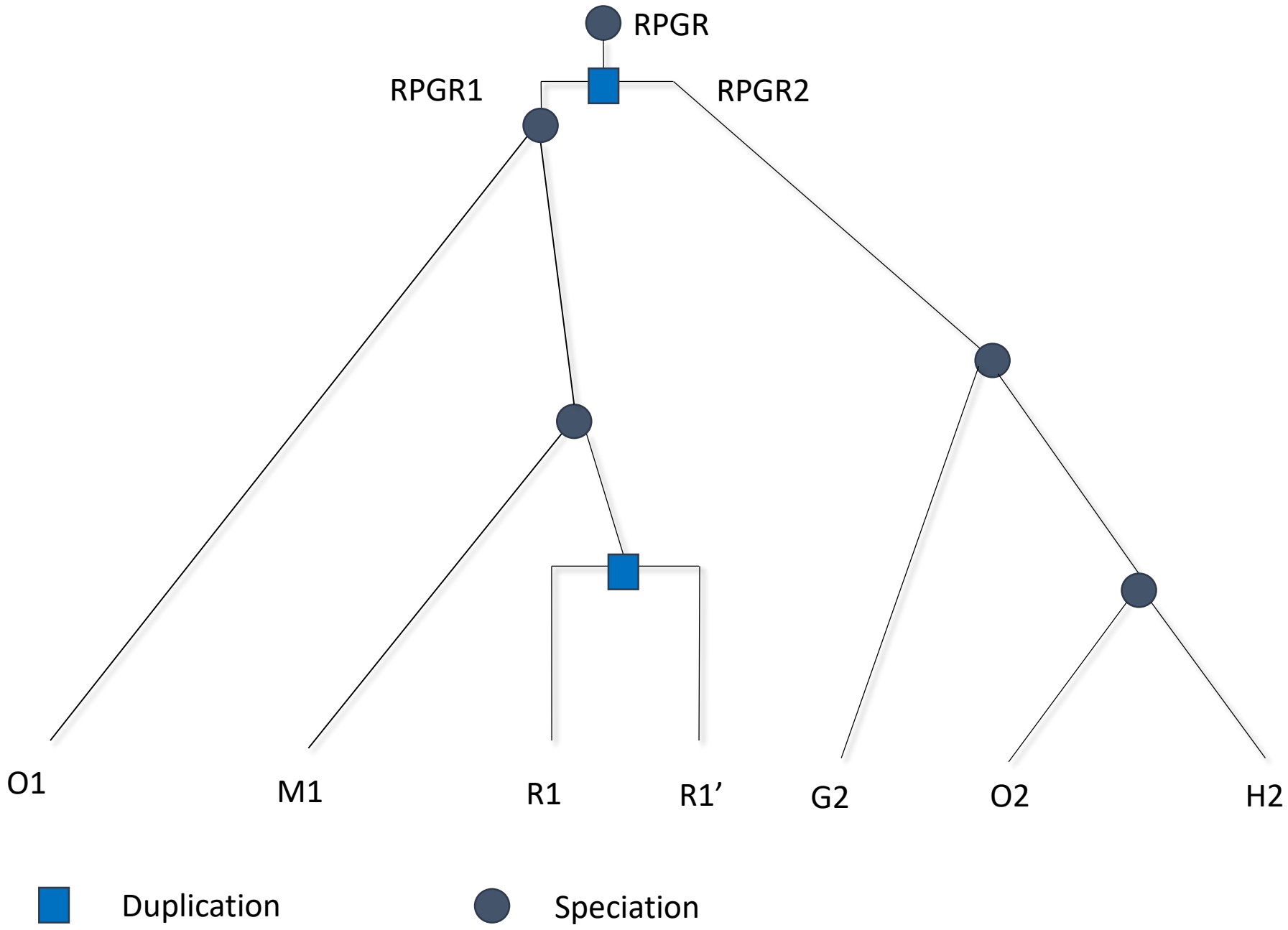

RPGR
Super-mammal
RPGR1
RPGR2
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

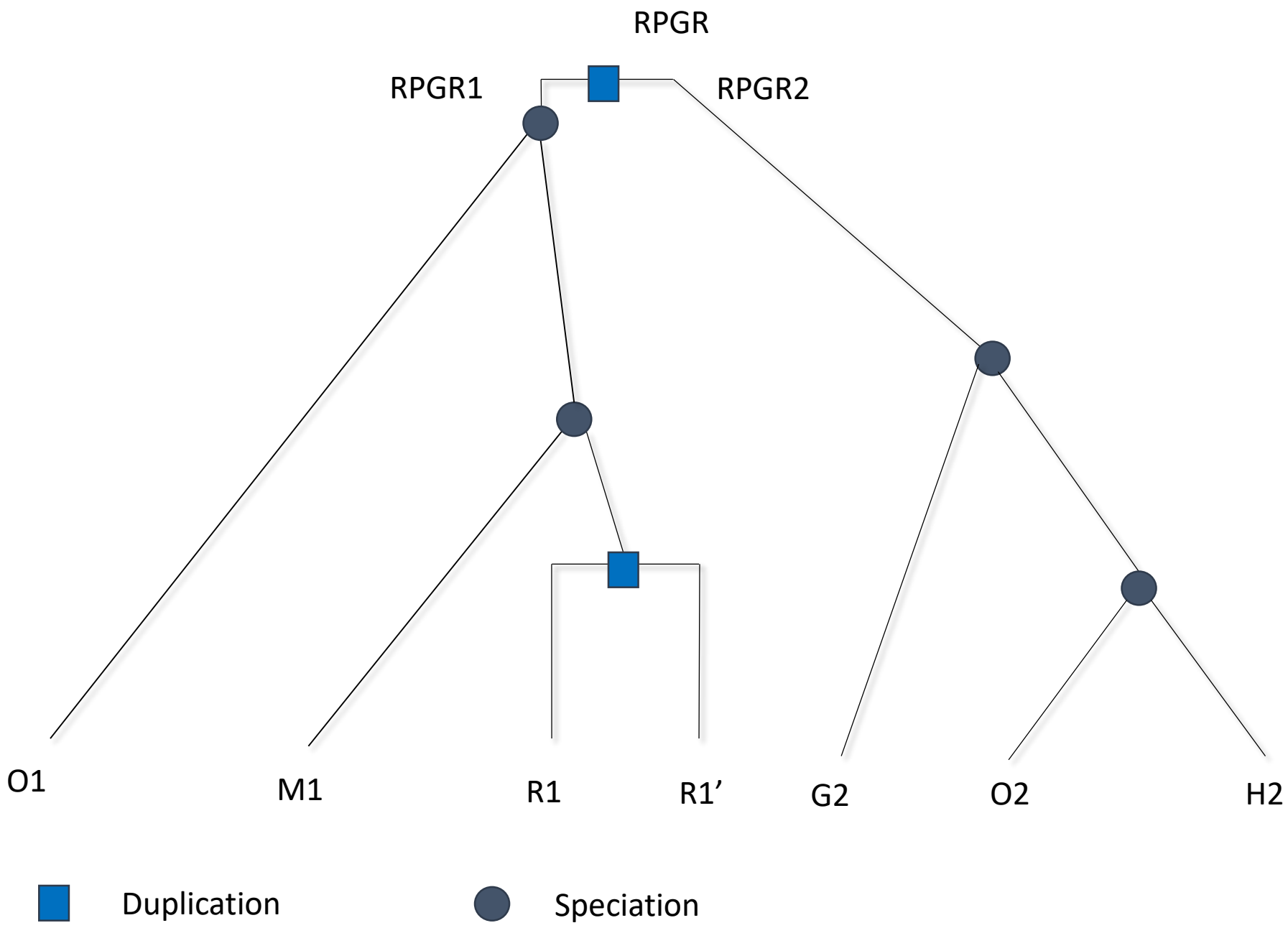
Notation tip: genes are labeled by their species.




RPGR
RPGR1
RPGR2
G2
O1
O2
H2
M1
R1
R1’
Duplication
Speciation


RPGR
RPGR1
RPGR2
O1
M1
R1
R1'
G2
O2
H2
Duplication
Speciation


RPGR
RPGR1
RPGR2
O1
M1
R1
R1'
G2
O2
H2
Duplication
Speciation


RPGR
RPGR1
RPGR2
O1
M1
R1
R1'
G2
O2
H2
Duplication
Speciation

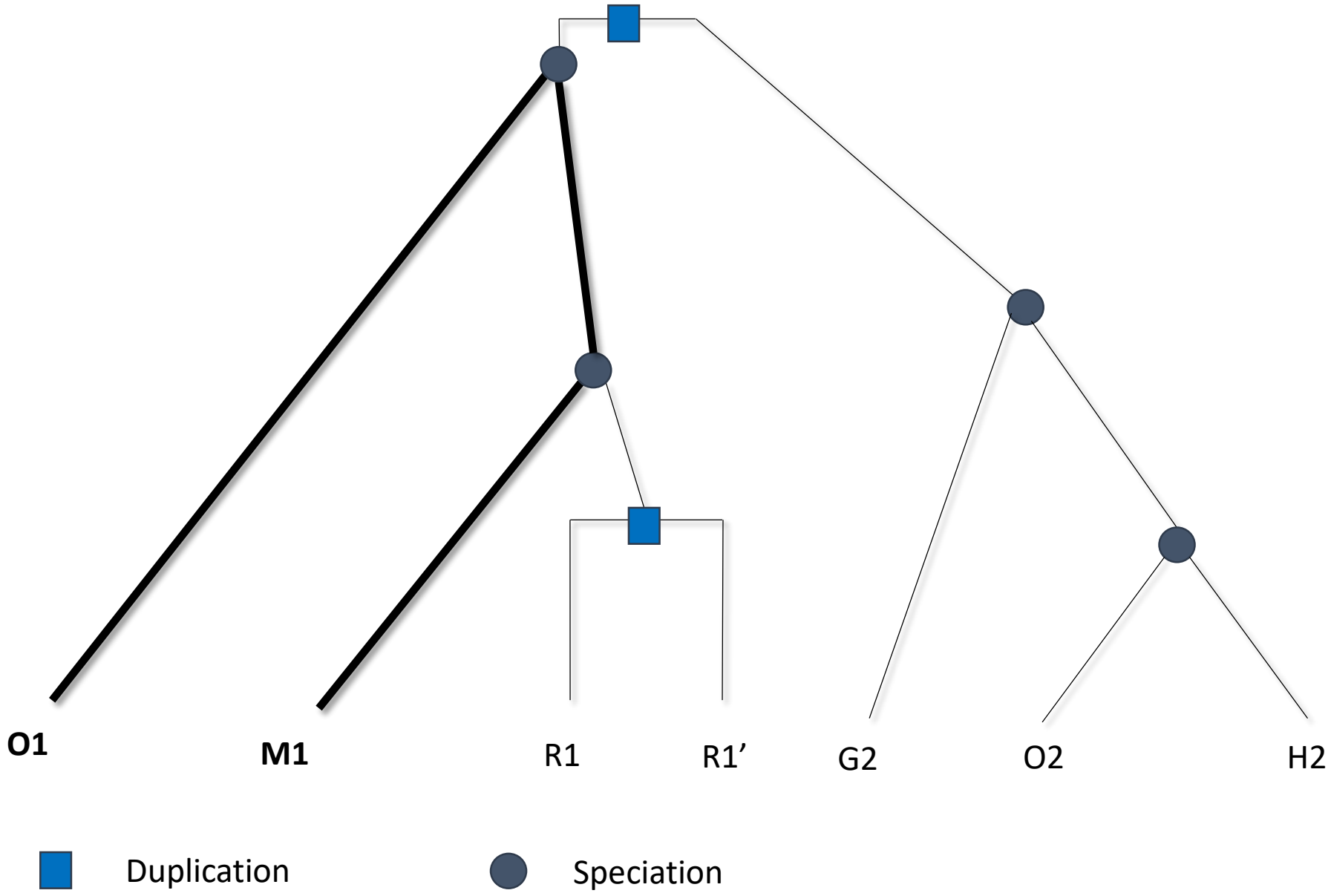
Orthologs and paralogs

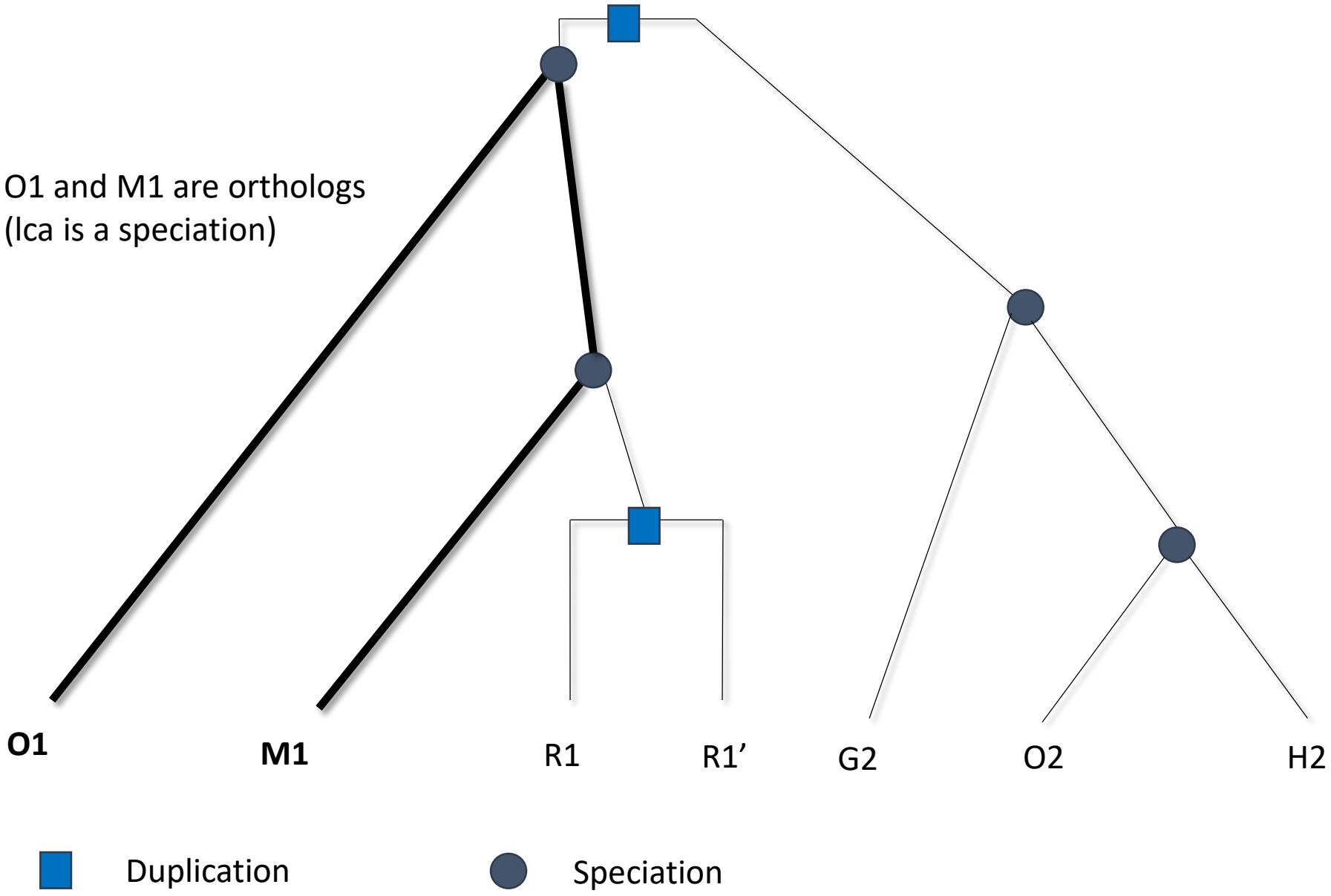
Two genes are*:

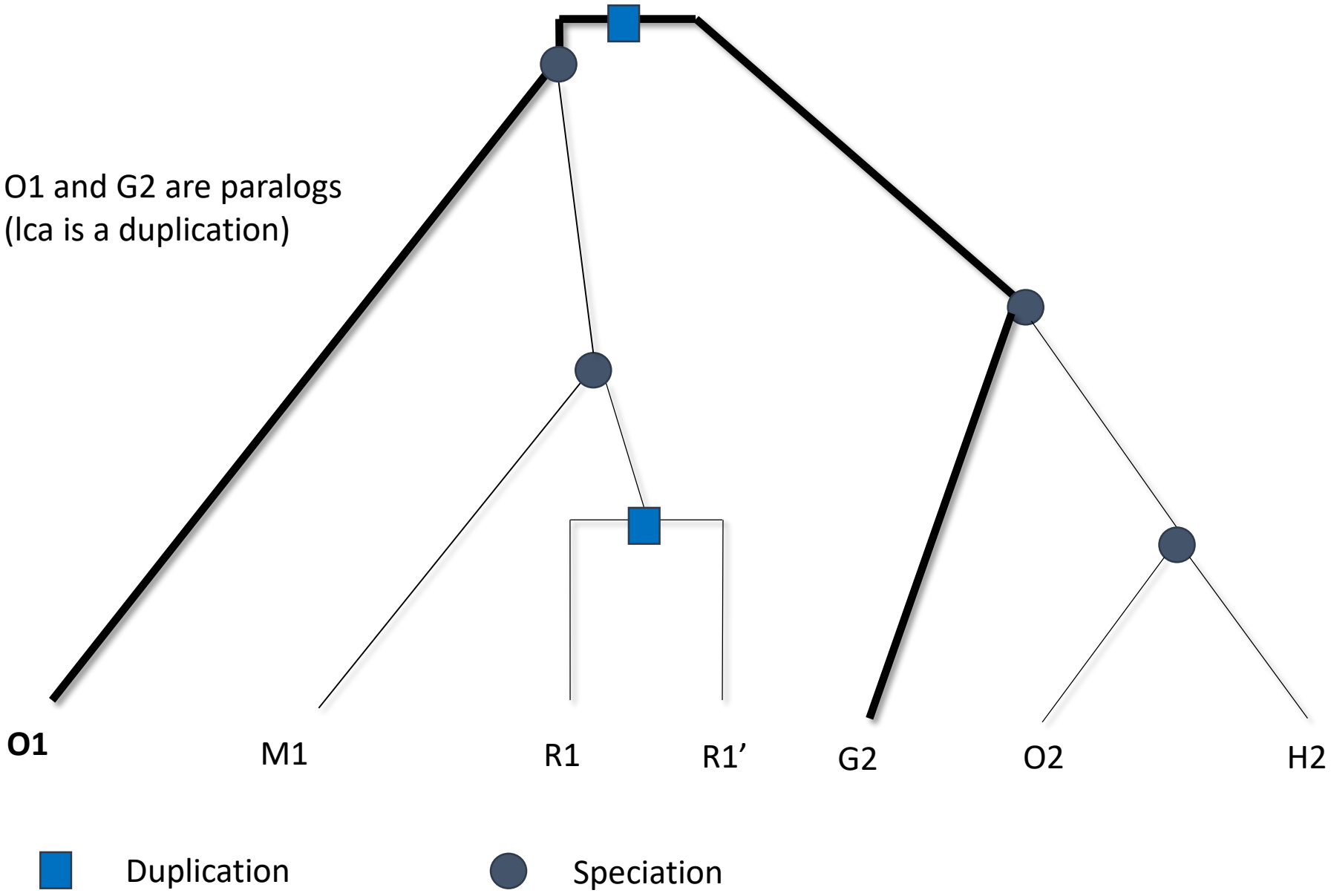
Orthologs if their lowest common ancestor underwent **speciation**

Paralogs if their lowest common ancestor underwent **duplication**

***w.r.t. a given gene tree**


O1
M1
R1
R1'
G2
O2
H2
Duplication
Speciation


O1 and M1 are orthologs
(lca is a speciation)
O1
M1
R1
R1'
G2
O2
H2
Duplication
Speciation


O1 and G2 are paralogs
(lca is a duplication)
O1
M1
R1
R1'
G2
O2
H2
Duplication
Speciation

Why bother?

Orthology/paralogy relations are related to gene functionality.

Some gene **functional annotation databases** assume that orthologs share the same functionality.

(e.g. COG, eggNOG databases)

Why bother?

Orthologs conjecture: orthologous genes tend to be similar in function, whereas paralogous genes tend to differ.

Why bother?

Orthologs conjecture: orthologous genes tend to be similar in function, whereas paralogous genes tend to differ.

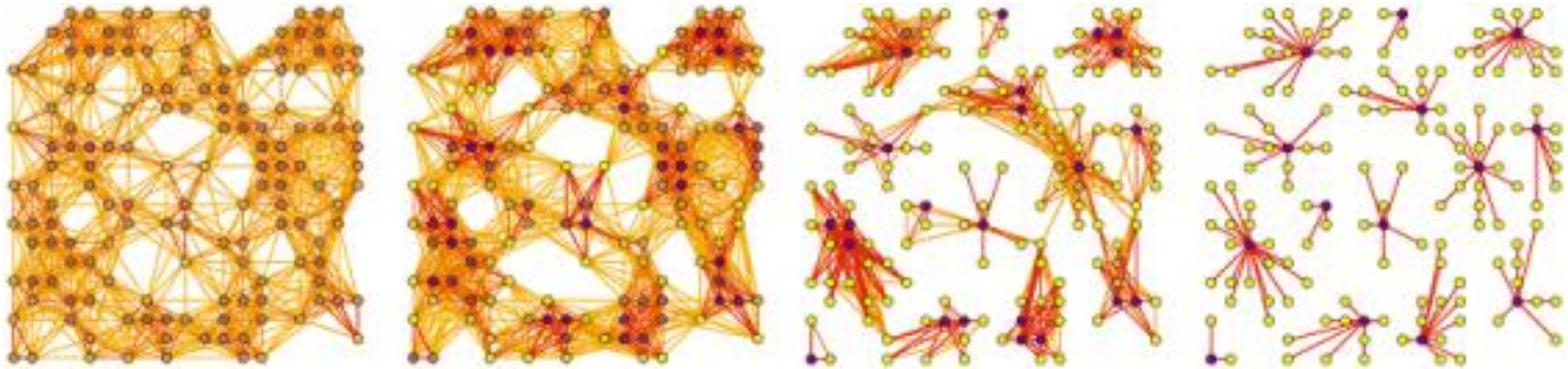
Quest For Orthologs consortium: "a joint effort to benchmark, improve and standardize orthology predictions through collaboration, the use of shared reference datasets, and evaluation of emerging new methods".

Traditional inference method

Clustering genes into groups of orthologs:

- If g1 and g2 and "**similar enough**" in terms of sequence, we say that g1 and g2 are putative orthologs.
- Make a graph G of putative orthologs.
- Partition G into clusters, i.e. highly connected components

  Otherwise, too many false positives occur
- OrthoMCL, InParanoid, proteinortho, …

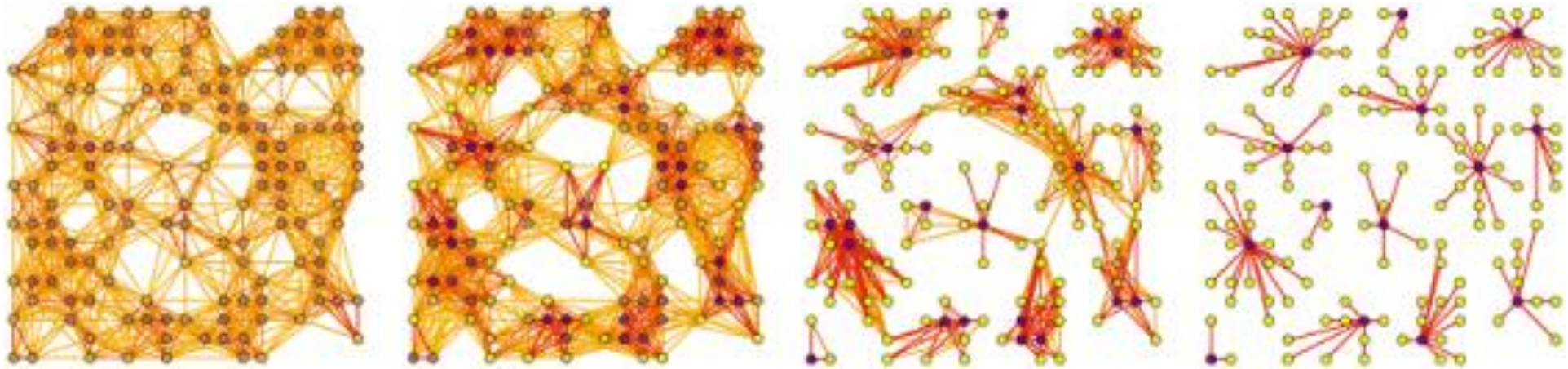


Traditional inference method

Clustering genes into groups of orthologs:

- If g1 and g2 and "**similar enough**" in terms of sequence, we say that g1 and g2 are putative orthologs
- Make a graph G of putative orthologs.
- Partition G into clusters, i.e. highly connected components

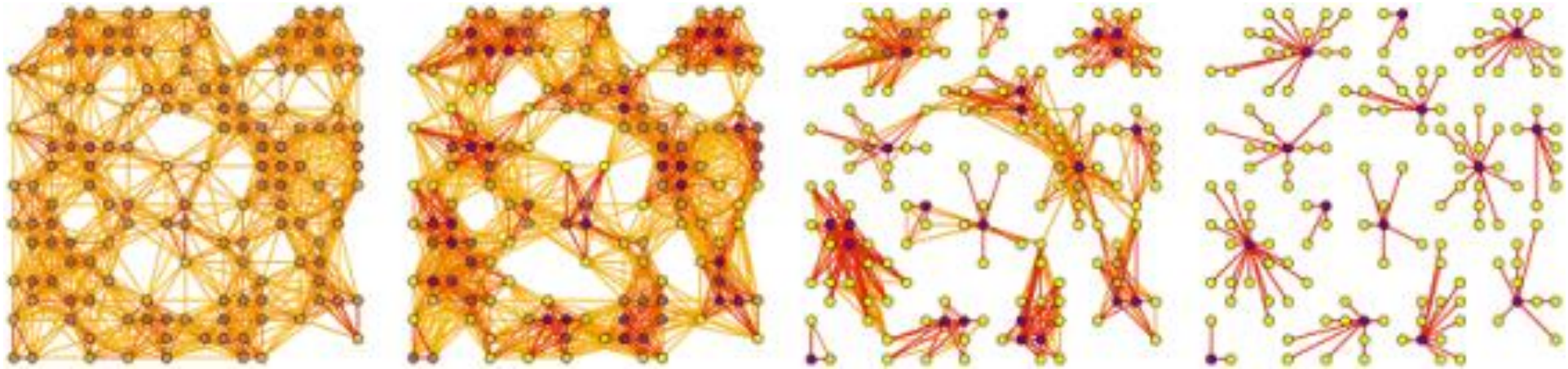
  Otherwise, too many false positives occur
- OrthoMCL, InParanoid, proteinortho, …



Traditional inference method

Clustering genes into groups of orthologs:

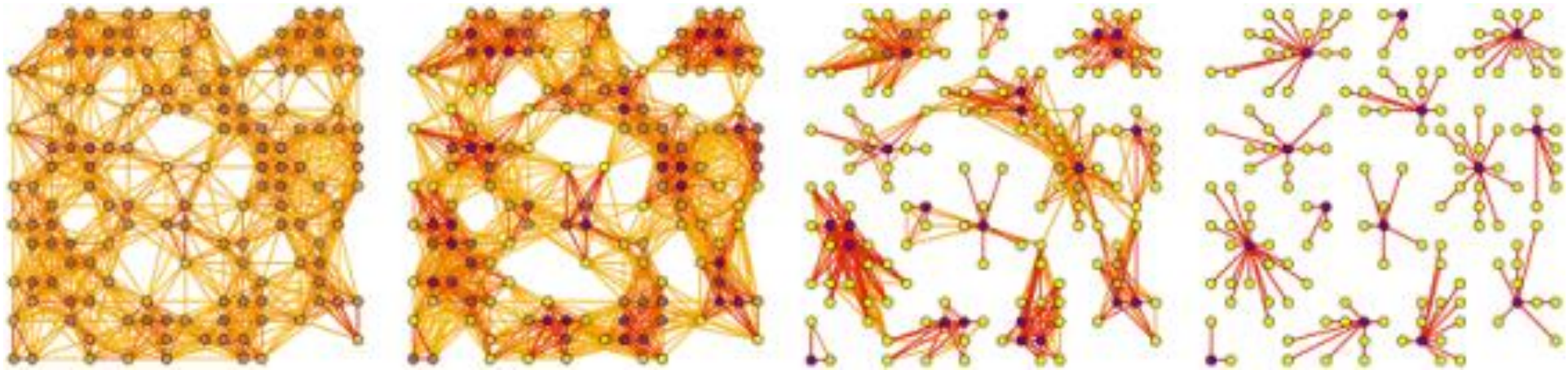
- If g1 and g2 and "**similar enough**" in terms of sequence, we say that g1 and g2 are putative orthologs
- "Similar enough" usually means that, if g1 and g2 are from species s1 and s2, they for a Bidirectional Best Hit (BBH):
  - g1's best match in s2 is g2
  - g2's best match in s1 is g1



Traditional inference method

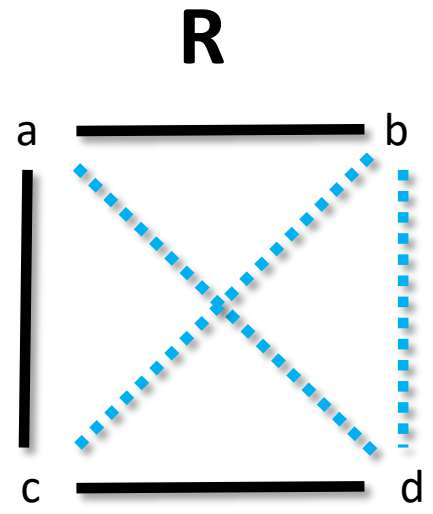
These methods are very often **incomplete** - have false positives or false negatives (according to our definitions).

In (Lafond & El-Mabrouk, 2014), we found that >70% of inferred sets of relations were **unsatisfiable** – corresponded to no possible gene tree.



Orthology/paralogy relation graph R

Sequences and stuff



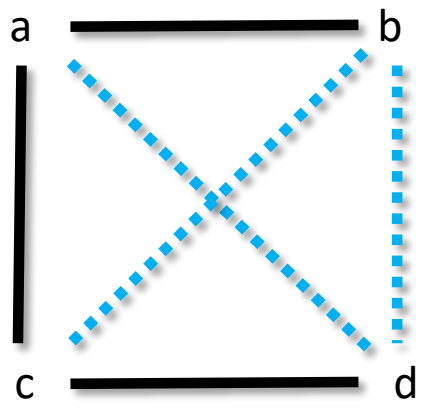
Orthologs = (a,b) (a, c) (c, d)

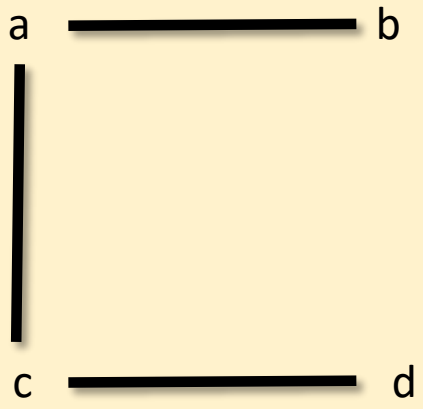
Paralogs = (a, d) (b, c) (b, d)

Orthology/paralogy graph

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

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





Notation tip:
sometimes, without warning,
edge = orthologs
non-edge = paralogs

Orthologs

Paralogs

What we want to do

Given a set of orthologs / paralogs in form of a relation graph R:

- Verify that they "**make sense**"

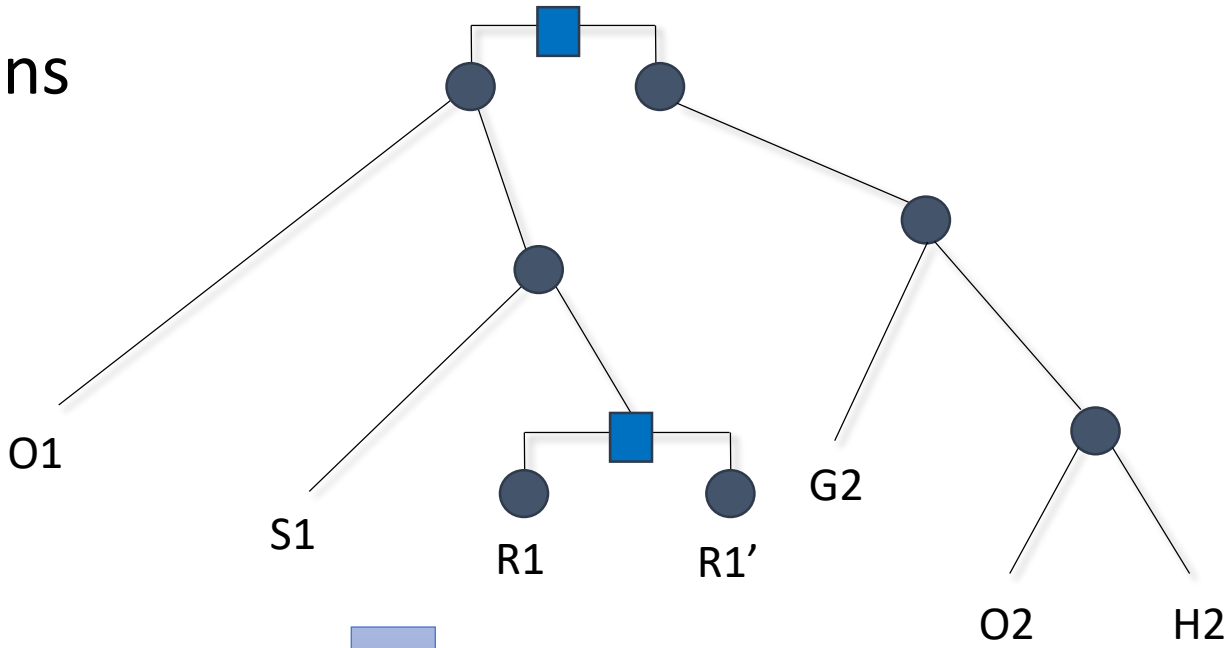
  Satisfiable: can some gene tree display the relations?

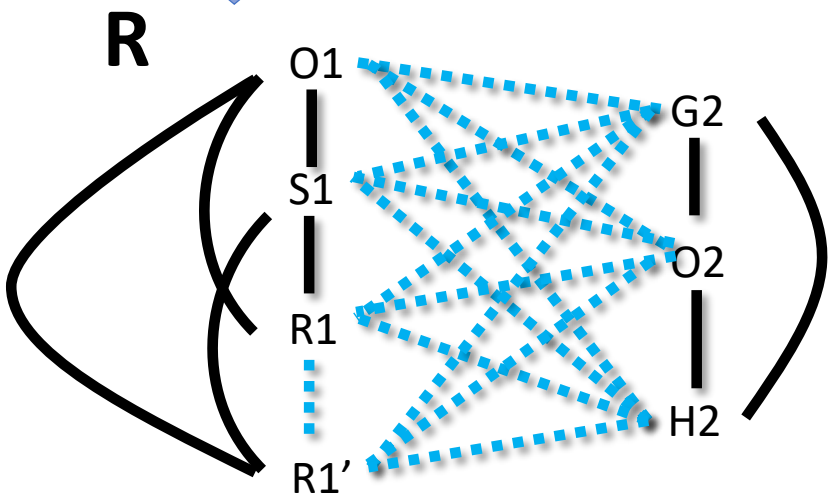
  Consistent: does it agree with our species tree?

- If they don't make sense, **correct them** in some minimal way

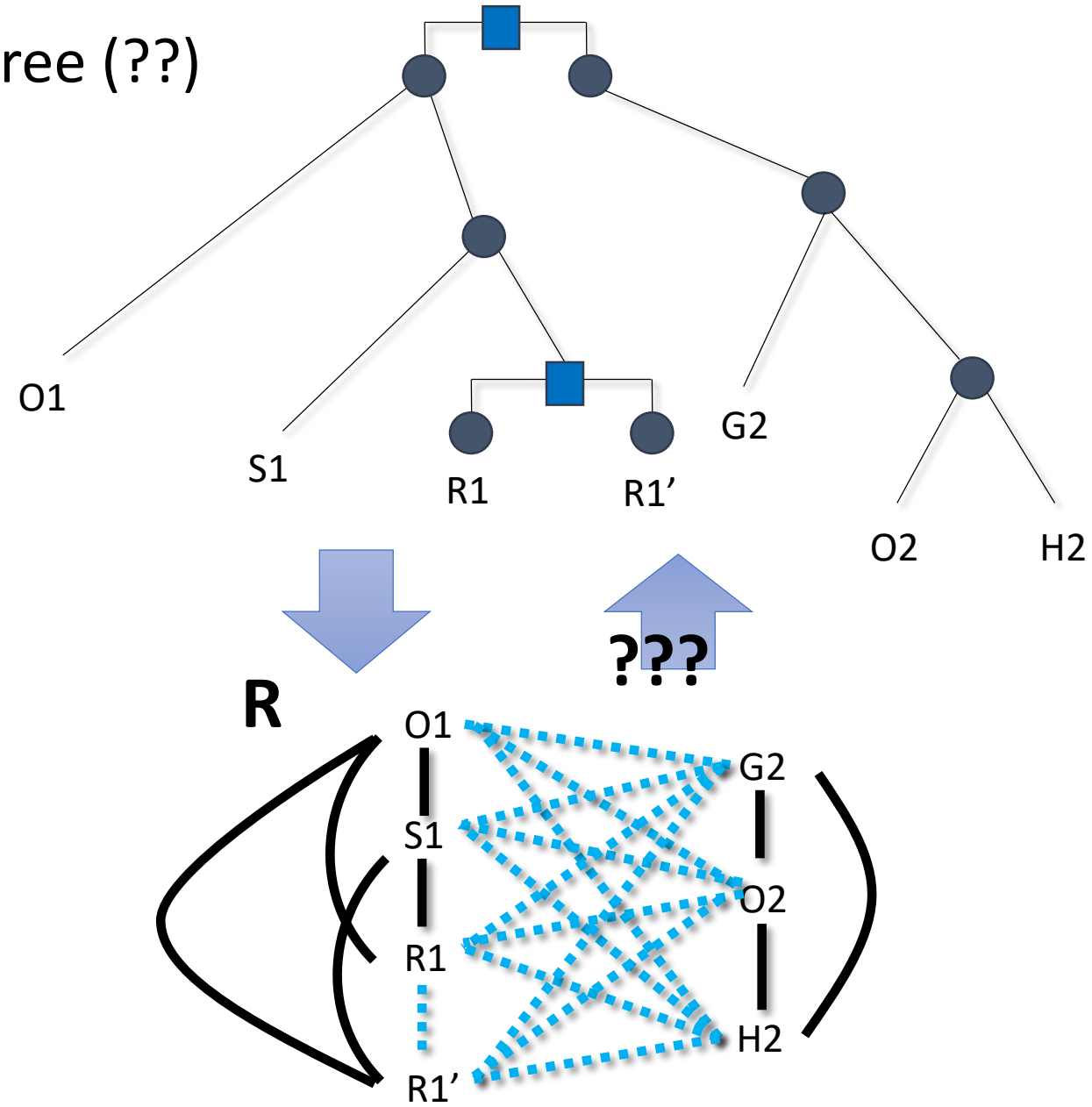
  Everything is NP-Complete

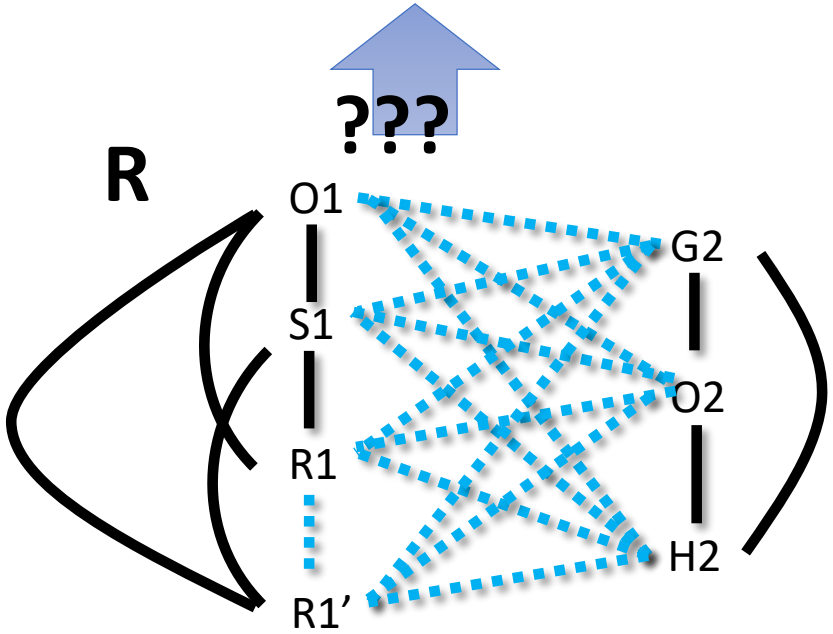
Gene tree => relations





Relations => Gene tree (??)

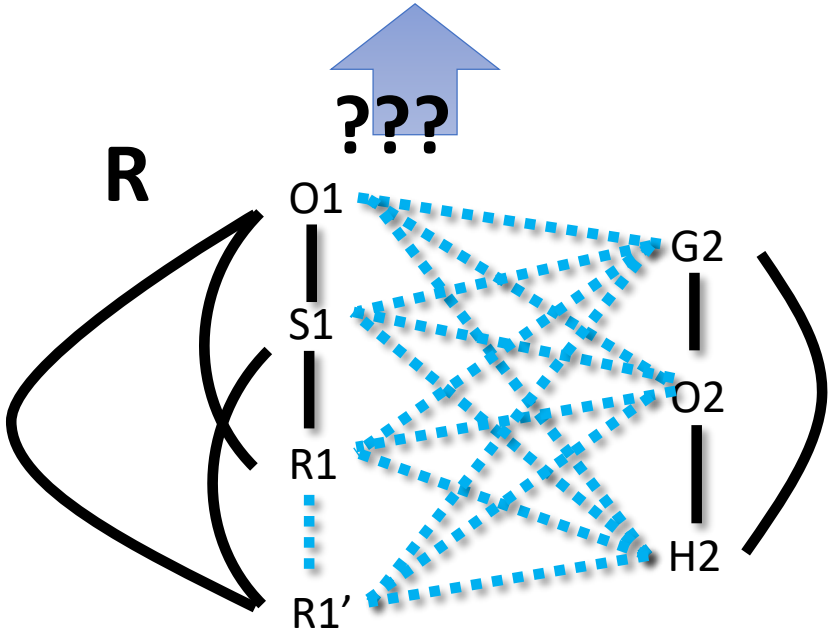



???
R
O1
S1
R1
R1'
G2
O2
H2

Problem :

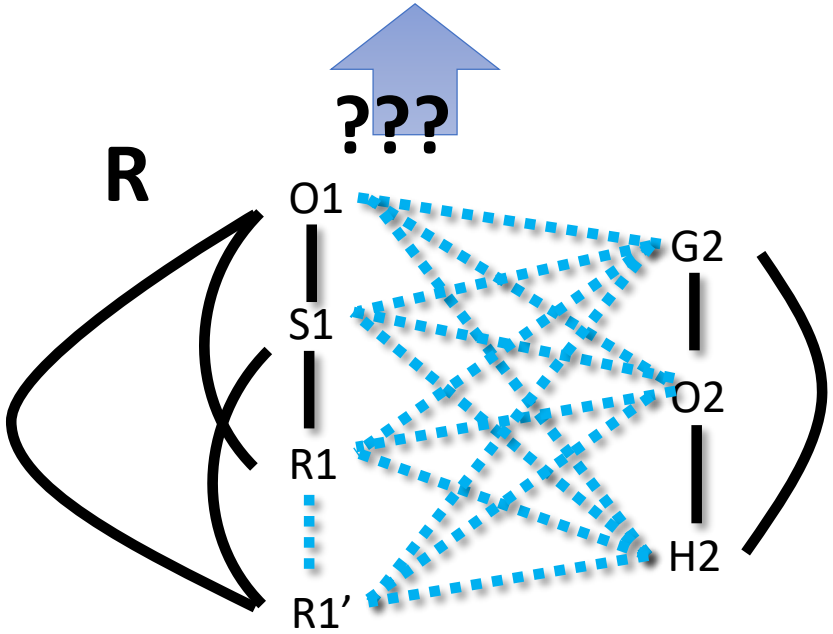
Given a relation graph R, is R **satisfiable**?

Does there exist a gene tree G that displays the relations of R ?



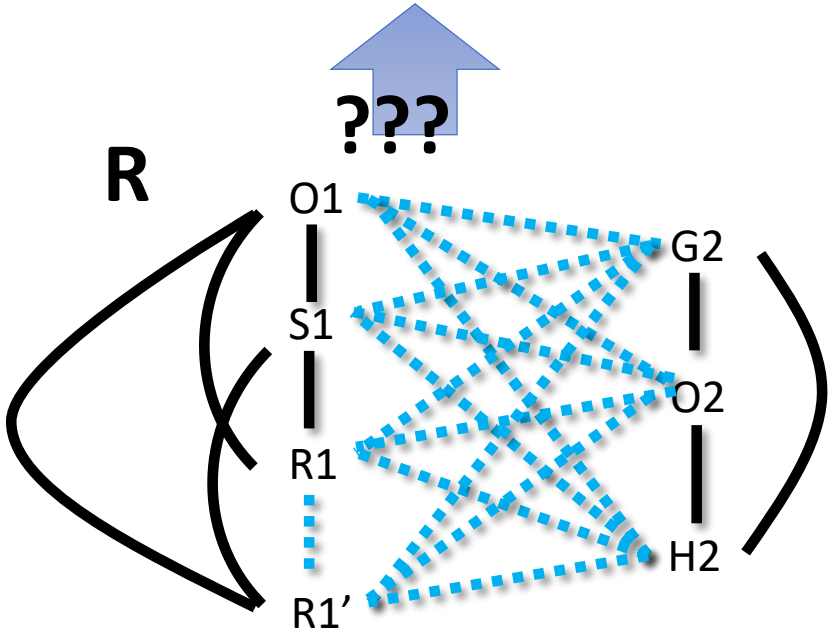
Usages of verifying satisfiability

1. Orthology graph benchmarking
2. Gene tree reconstruction
3. Species tree reconstruction



So, how do we verify whether there is a gene tree displaying these relations?

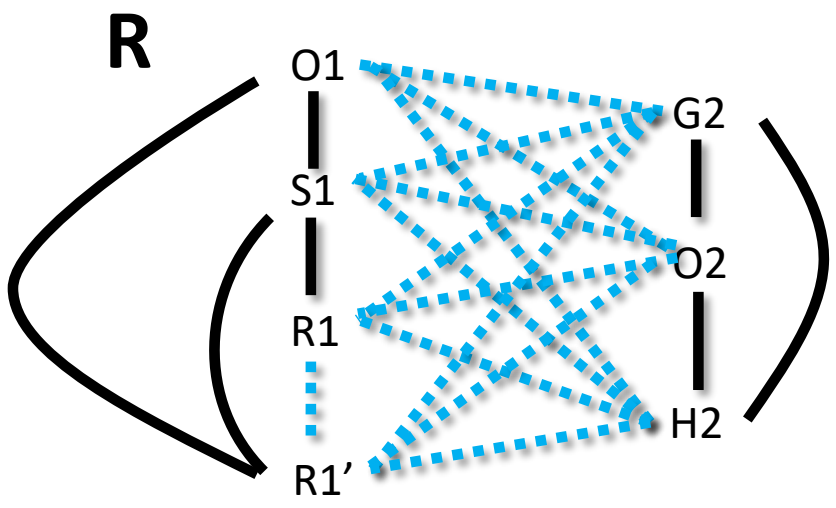
And if so, can we construct the tree?

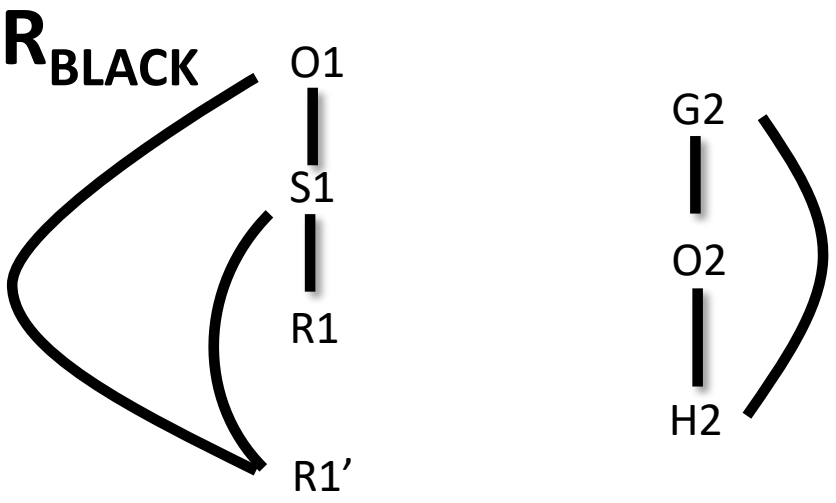


Theorem (Hernandez-Rosales & al., 2012):

A relation graph R is satisfiable if and only if R_{BLACK} is P_4-free (has no induced path on 4 vertices).

(P_4-free graphs are sometimes known as cographs)

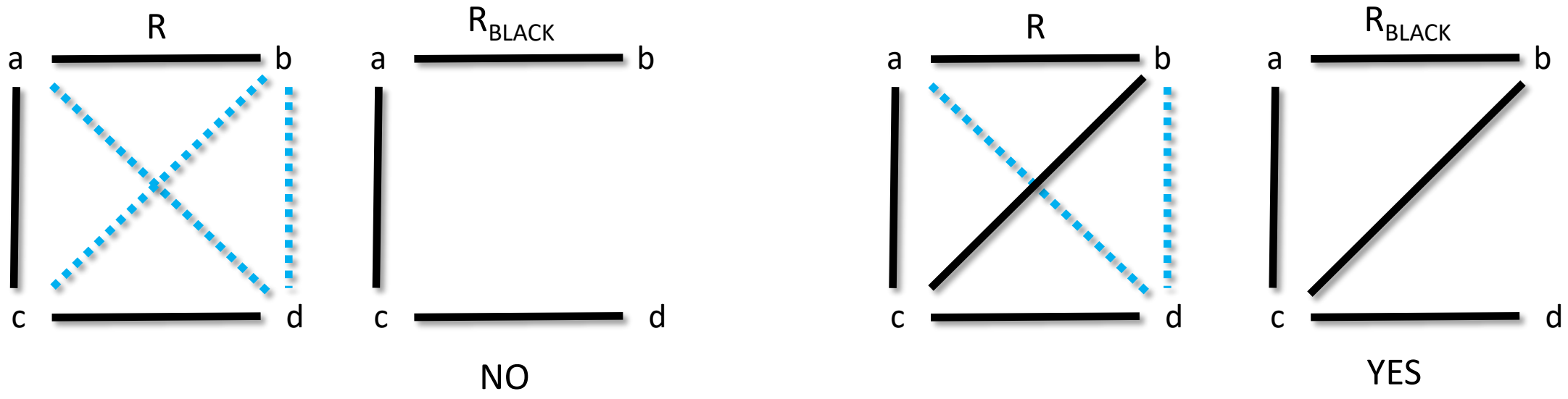




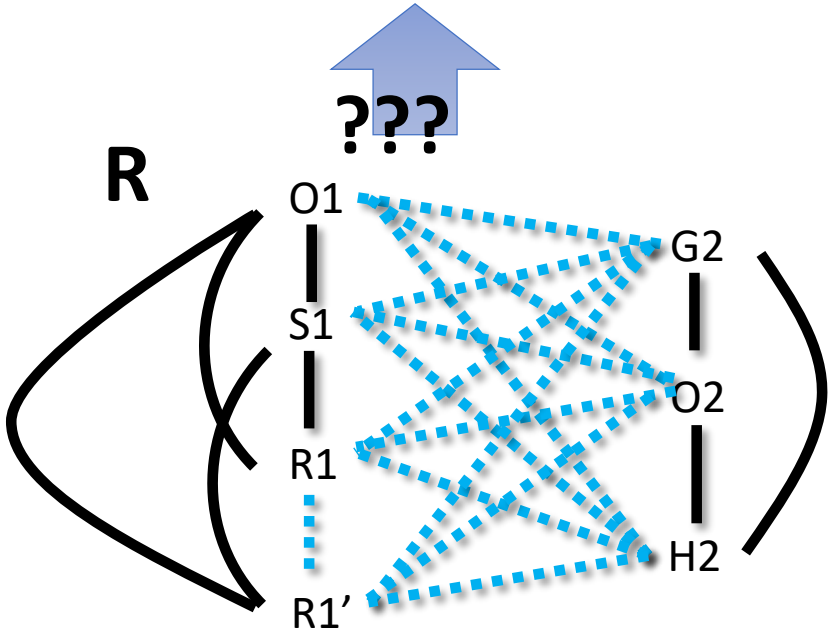
Theorem (Hernandez-Rosales & al., 2012):

A relation graph R is satisfiable if and only if R_{BLACK} is P_4-free (has no induced path on 4 vertices).

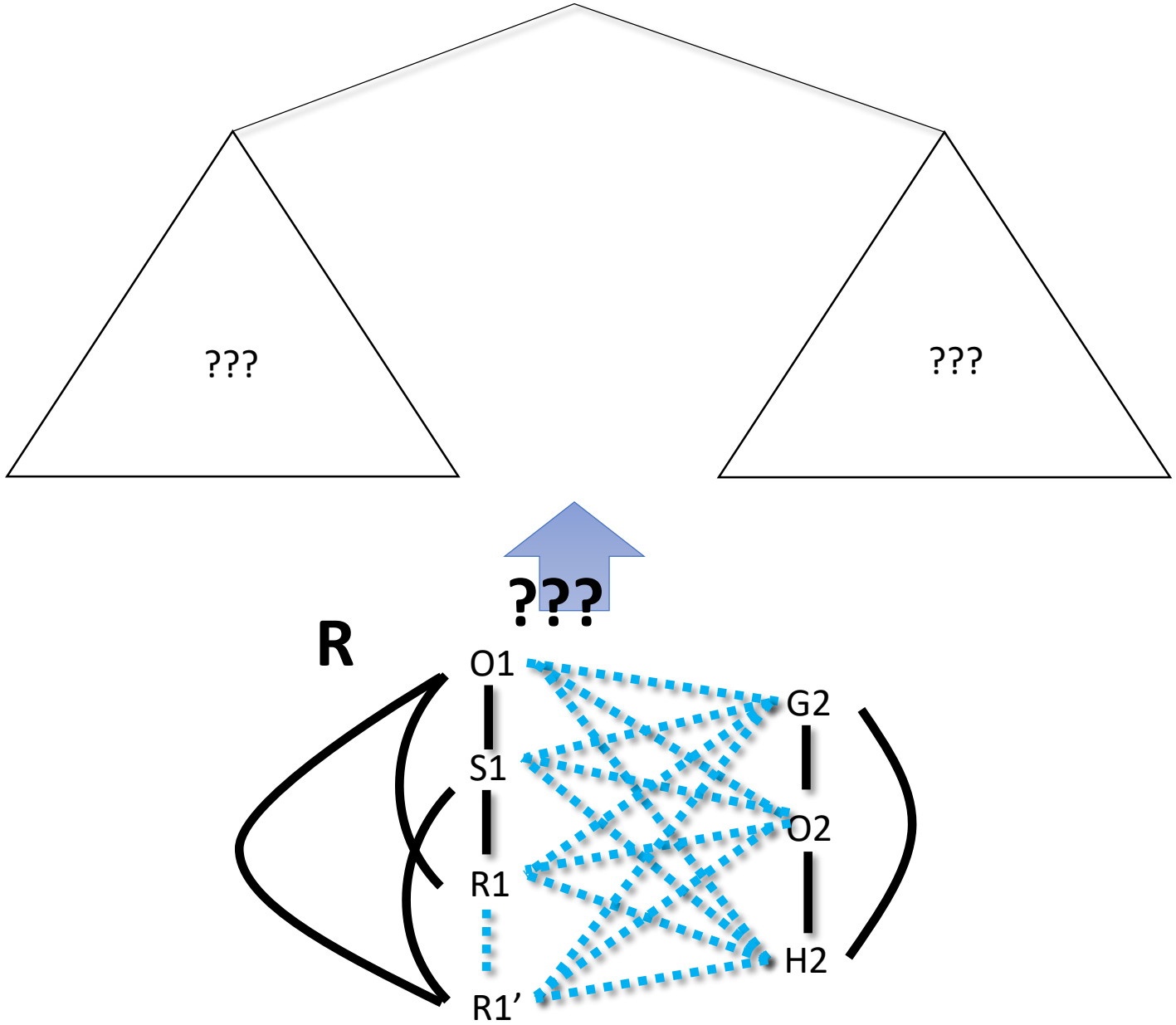
(P_4-free graphs are sometimes known as cographs)

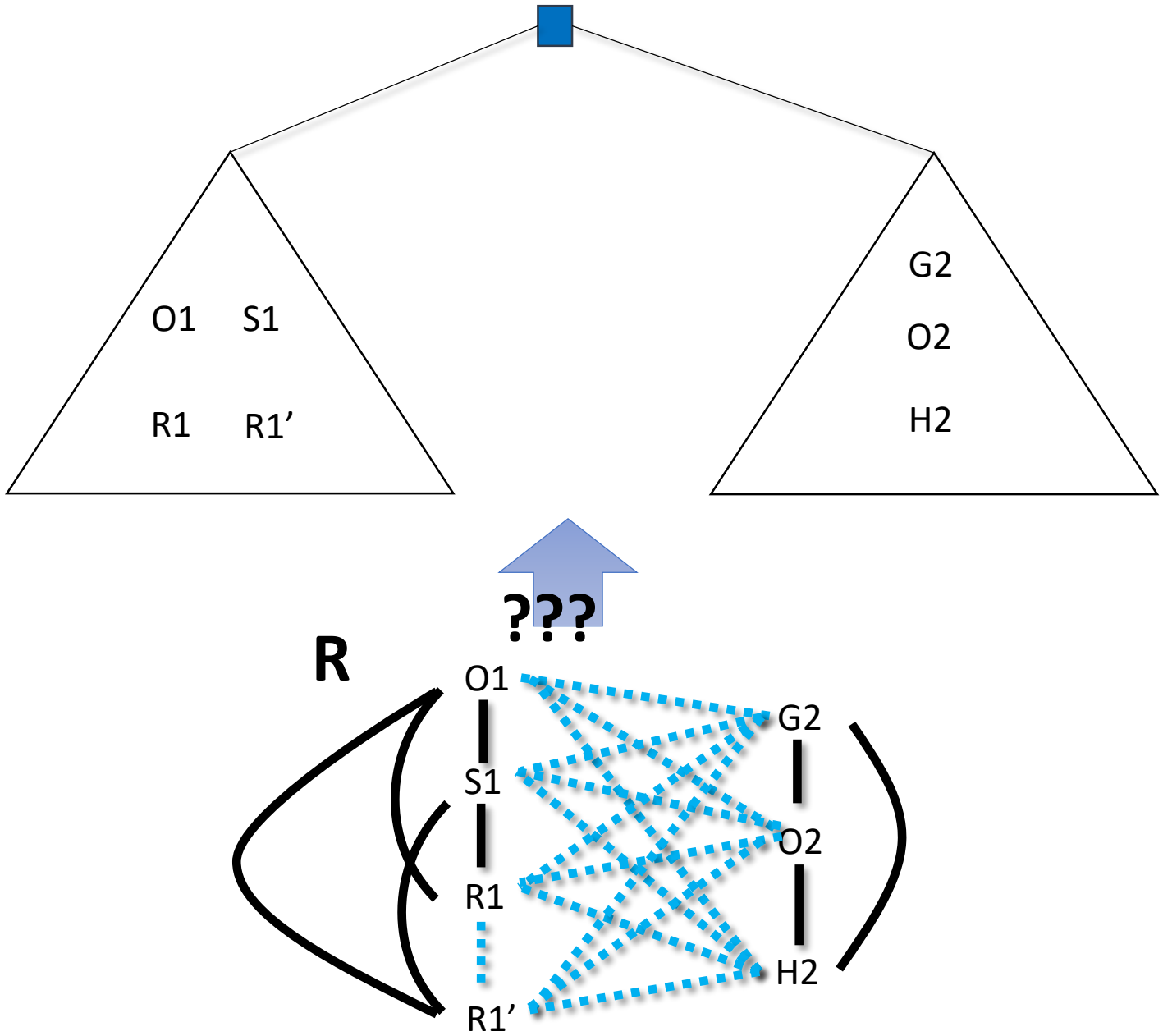


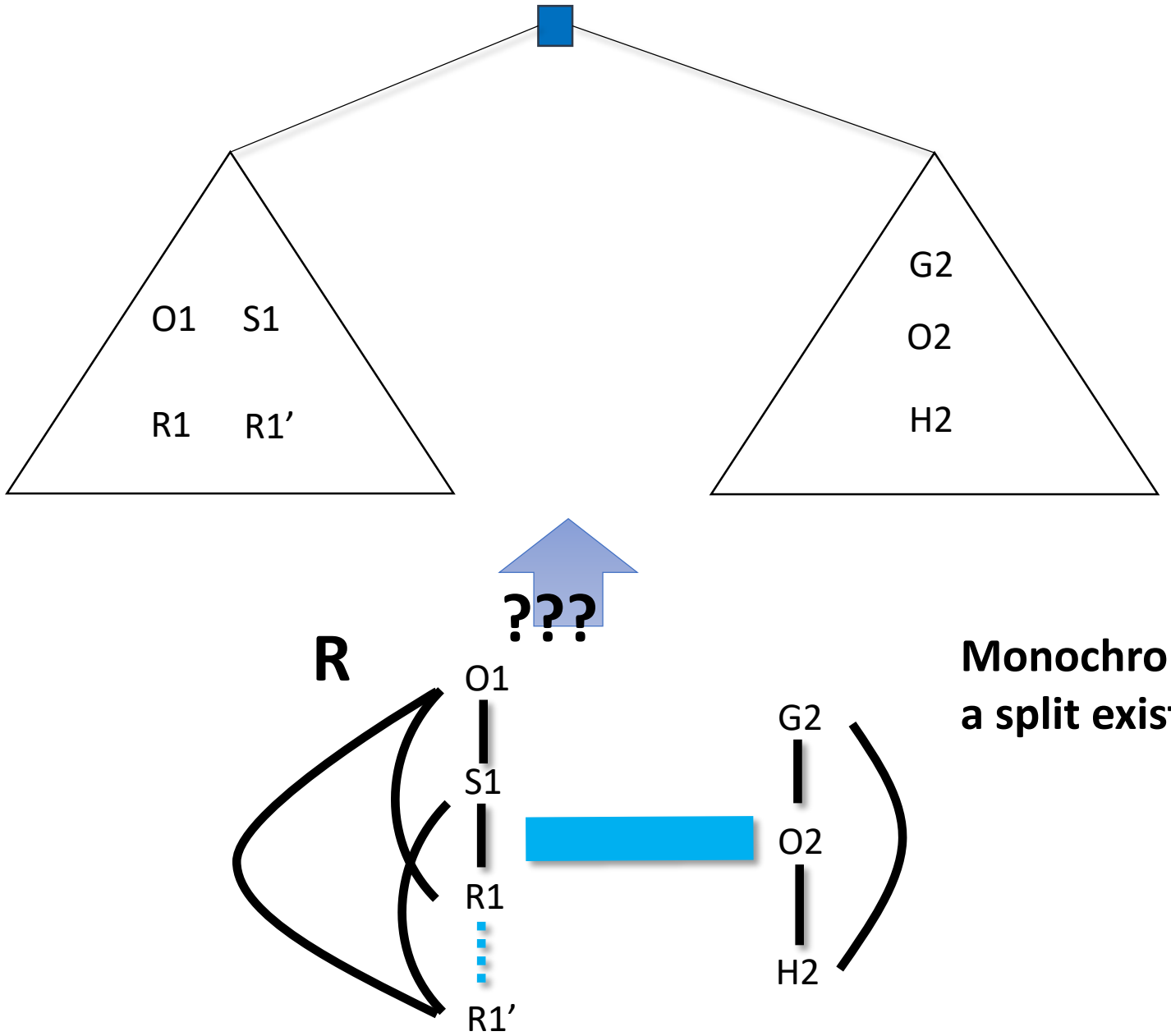
Is there a gene tree for R ?



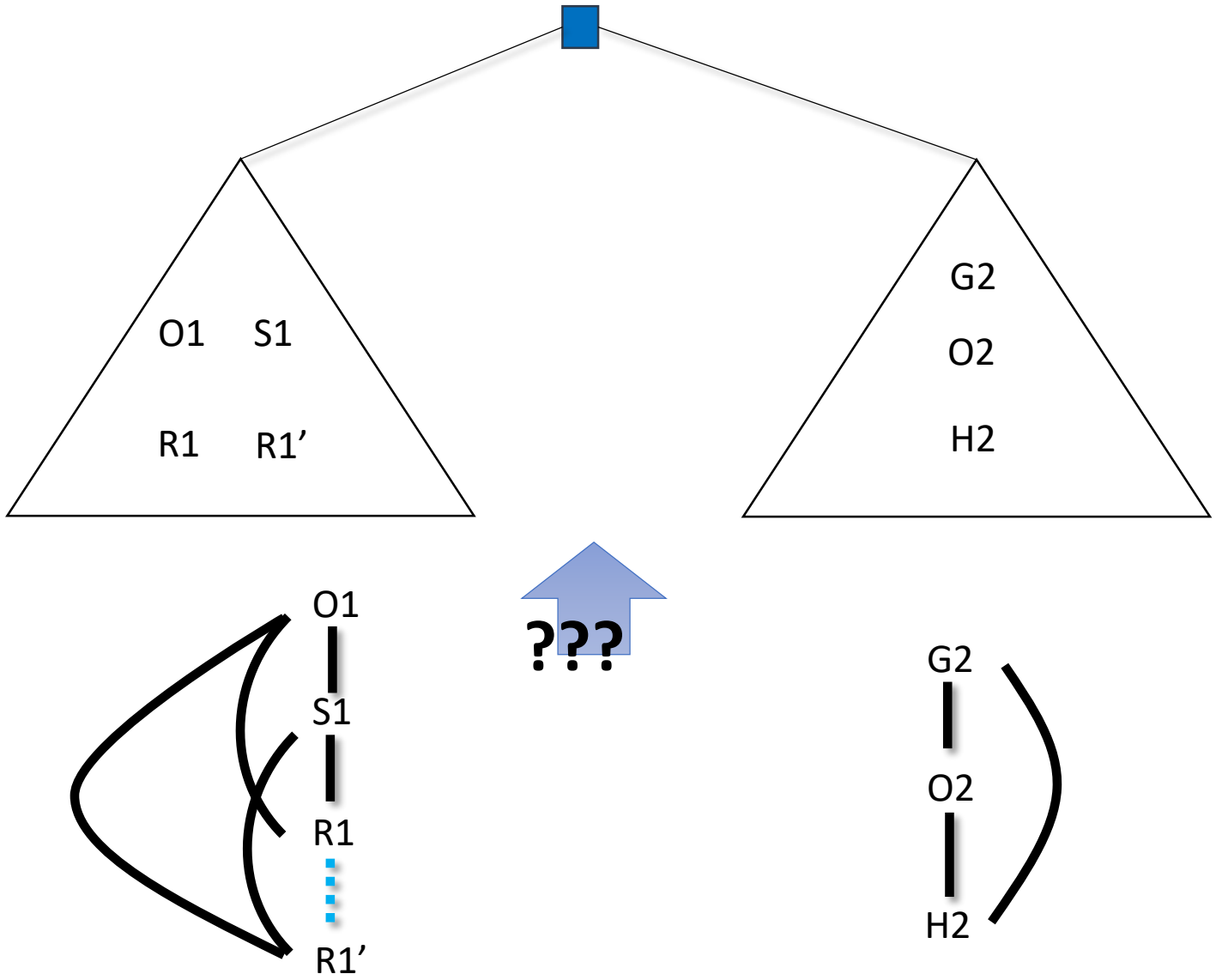
Let's say it exists…what is the first split then ?

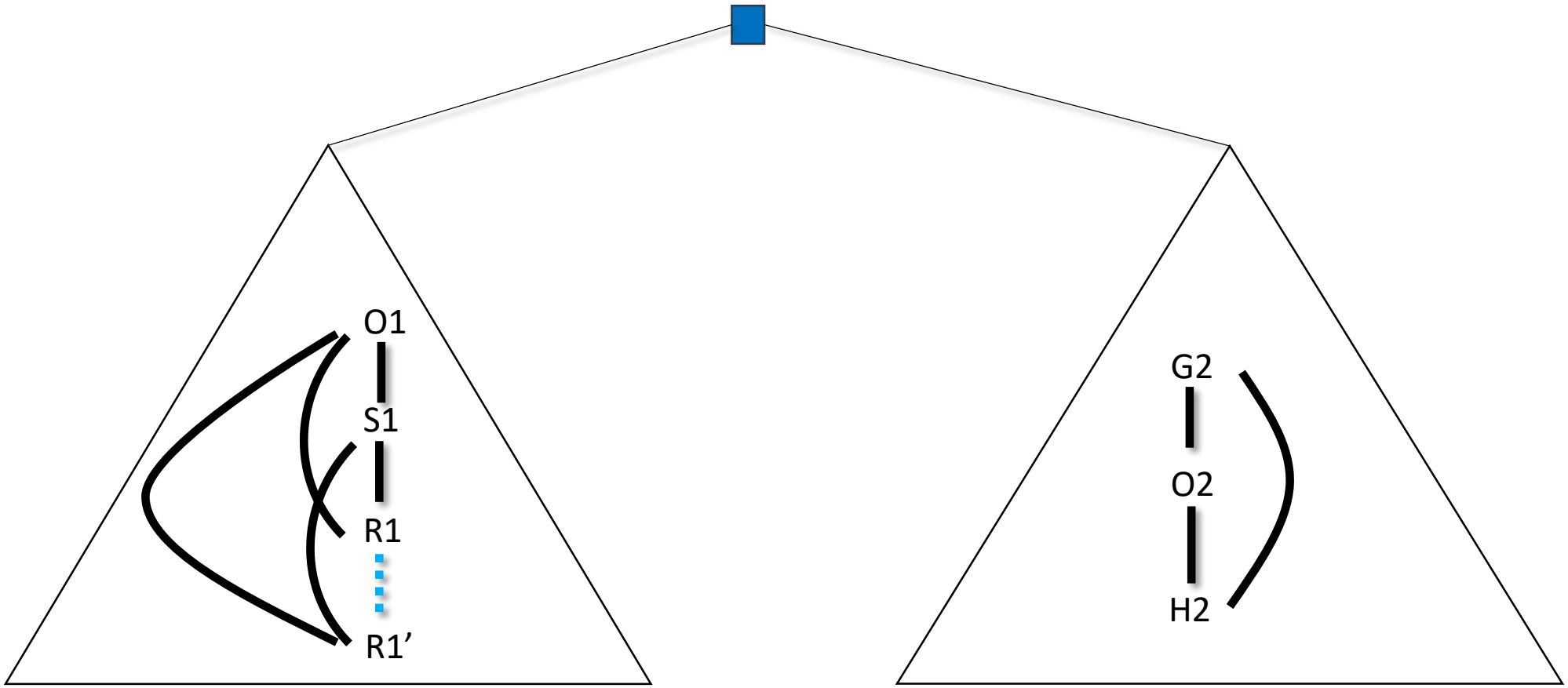


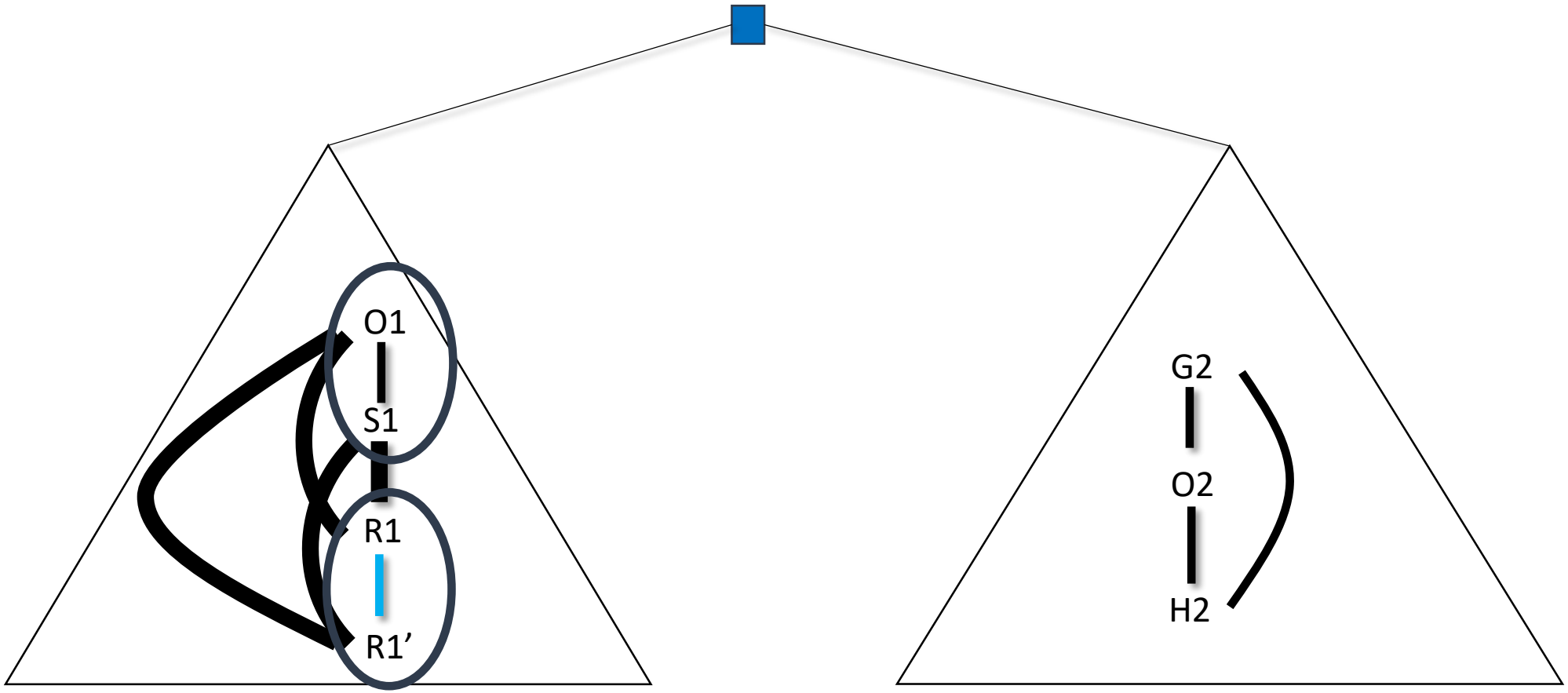

O1
S1
R1
R1'
G2
O2
H2
???
R
O1
S1
R1
R1'
G2
O2
H2

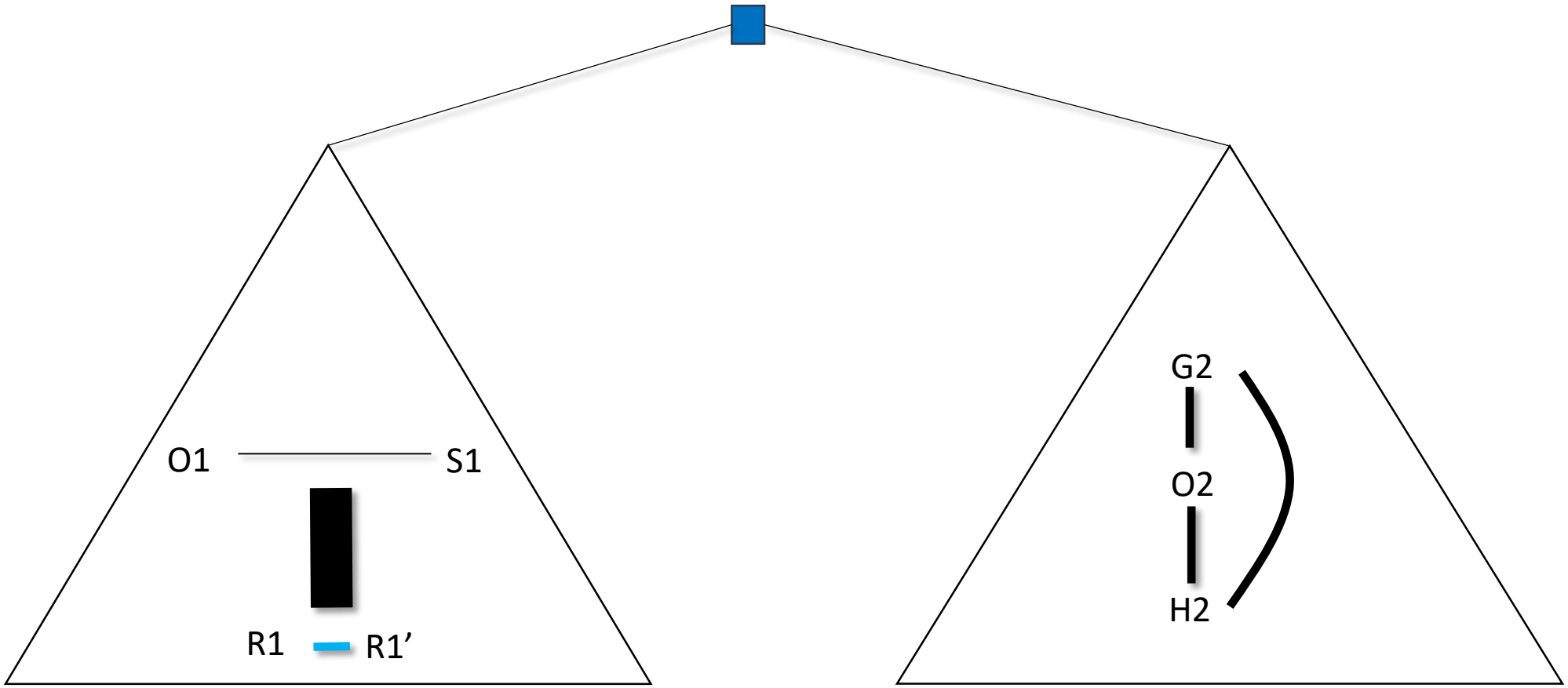


Monochromatic edge-cut => a split exists

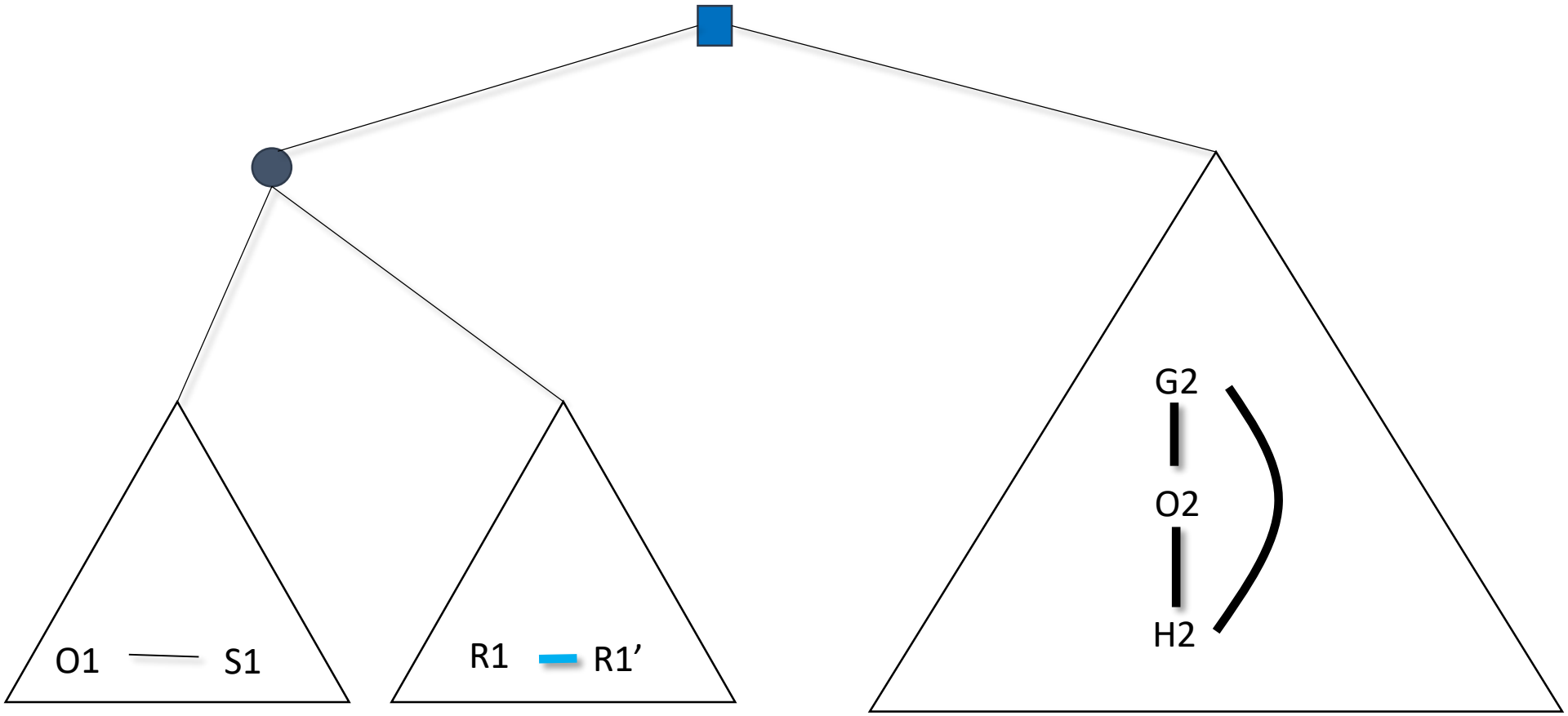

O1
S1
R1
R1'
G2
O2
H2
???
O1
S1
R1
R1'
G2
O2
H2


O1
S1
R1
R1’
G2
O2
H2


O1
S1
R1
R1’
G2
O2
H2



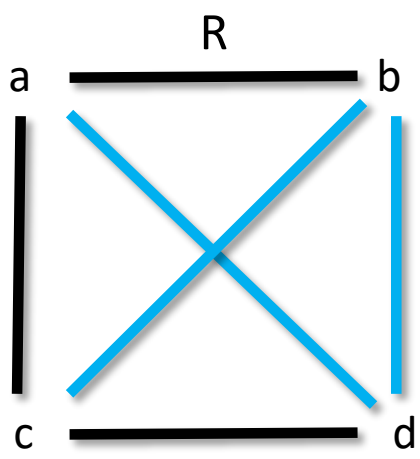
Monochromatic edge-cut => a split exists

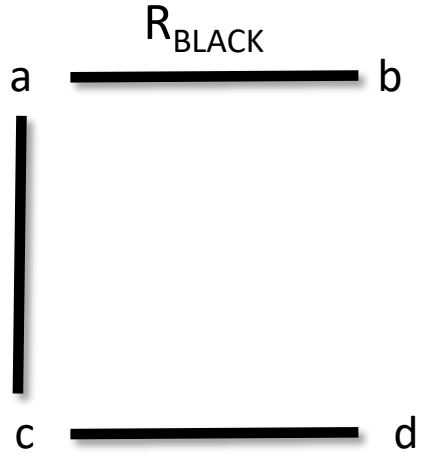


and so on …

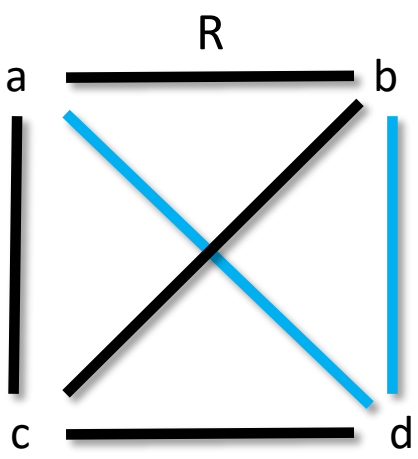
Theorem (informal) (Corneil, Perl & Stewart, 1985)

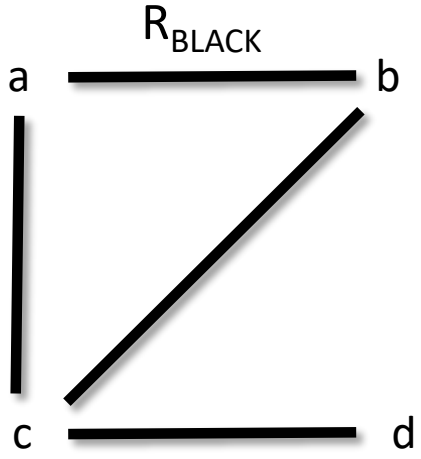
A monochromatic edge-cut will always exist if and only if R_{BLACK} is P_4-free.





NO





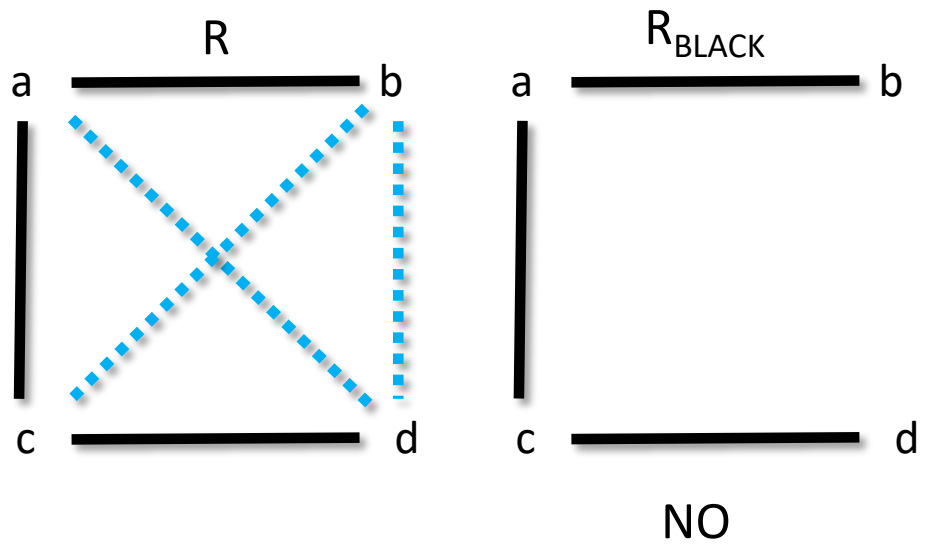
YES

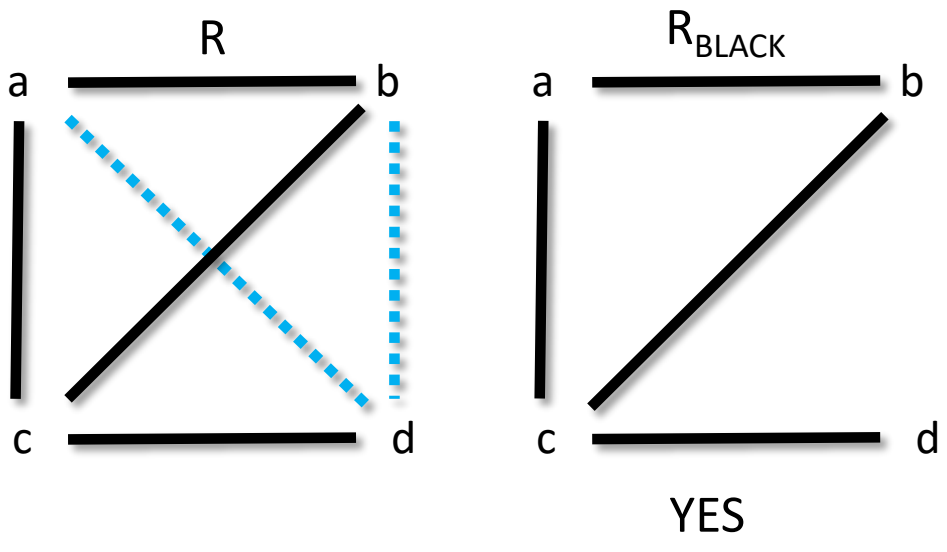
Theorem (informal) (Corneil, Perl & Stewart, 1985)

A monochromatic edge-cut will always exist if and only if R_{BLACK} is P_4-free.

P_4-freeness is easy to check in polynomial time.

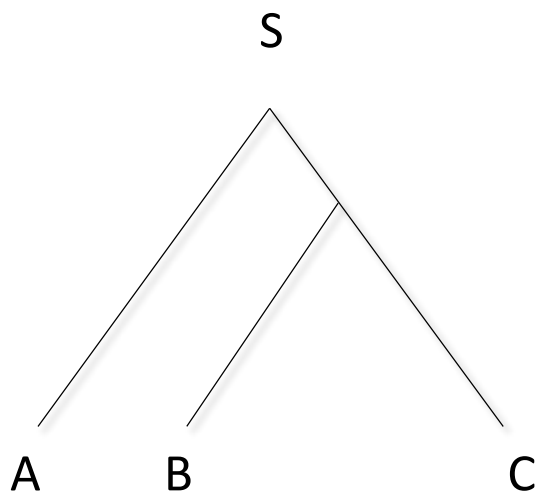
$O(n^4)$ in the obvious way, $O(n)$ in more clever ways.

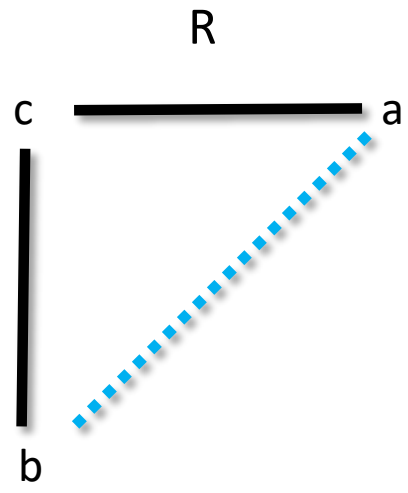




S-Consistency

What if we want our relations to agree with a given species tree S?

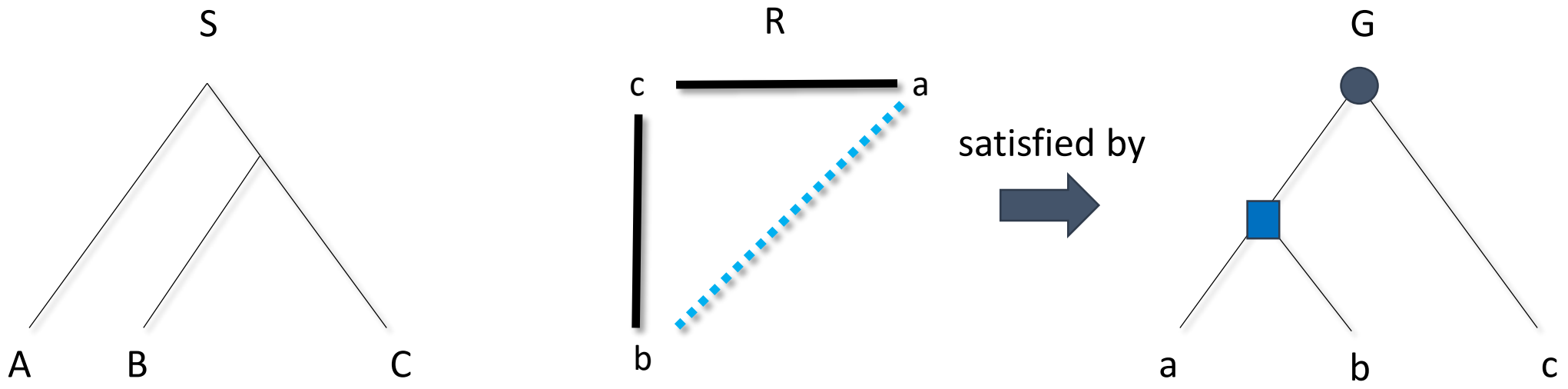




a = gene from species A
b = gene from species B
c = gene from species C

S-Consistency

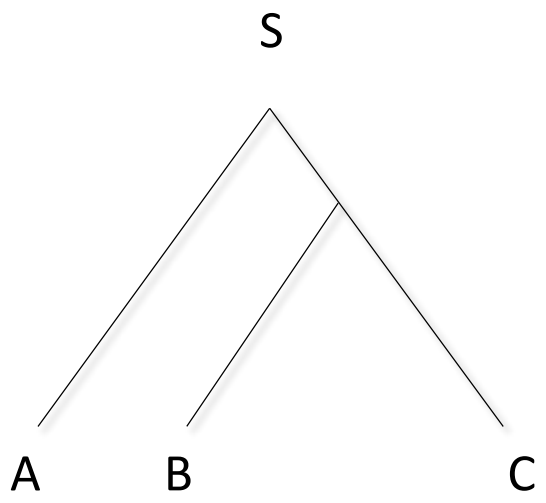
What if we want our relations to agree with a given species tree S?

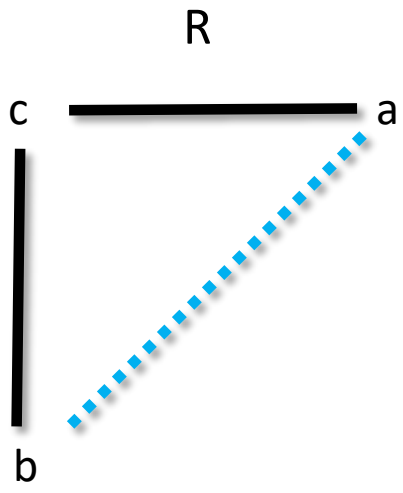


a = gene from species A
b = gene from species B
c = gene from species C

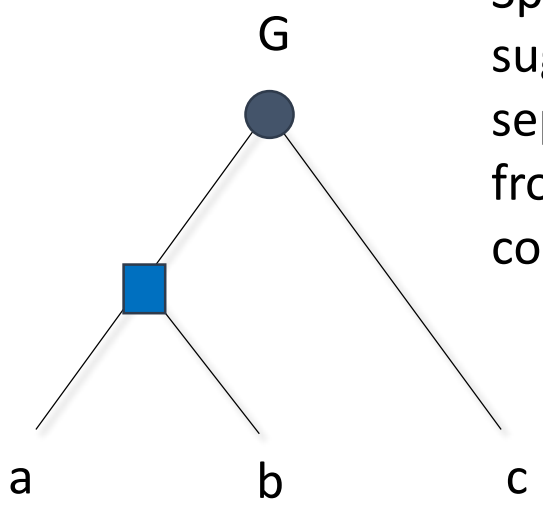
S-Consistency

What if we want our relations to agree with a given species tree S?





satisfied by



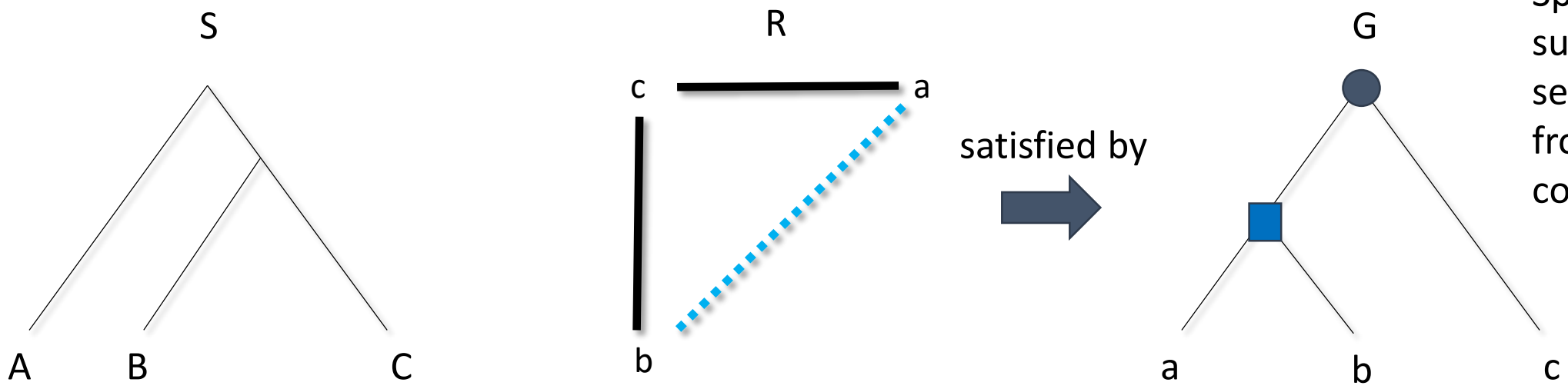
Speciation suggests separating (ab) from c, contradicting S

a = gene from species A
b = gene from species B
c = gene from species C

S-Consistency

What if we want our relations to agree with a given species tree S?

Can be checked in time $O(n^3)$ **(Hernandez-Rosales, 2012)**



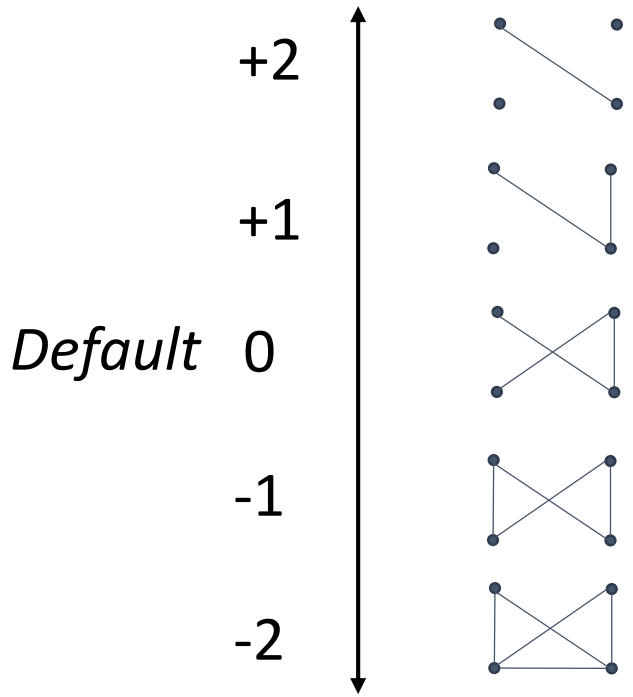
Speciation suggests separating (ab) from c, contradicting S

a = gene from species A
b = gene from species B
c = gene from species C

Experiments

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

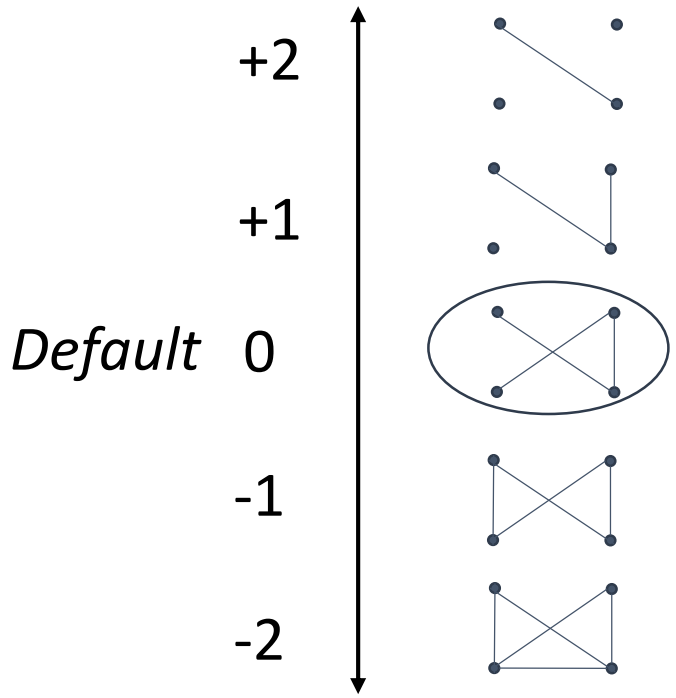
Stricter => Less orthologies



Looser => More orthologies

Experiments

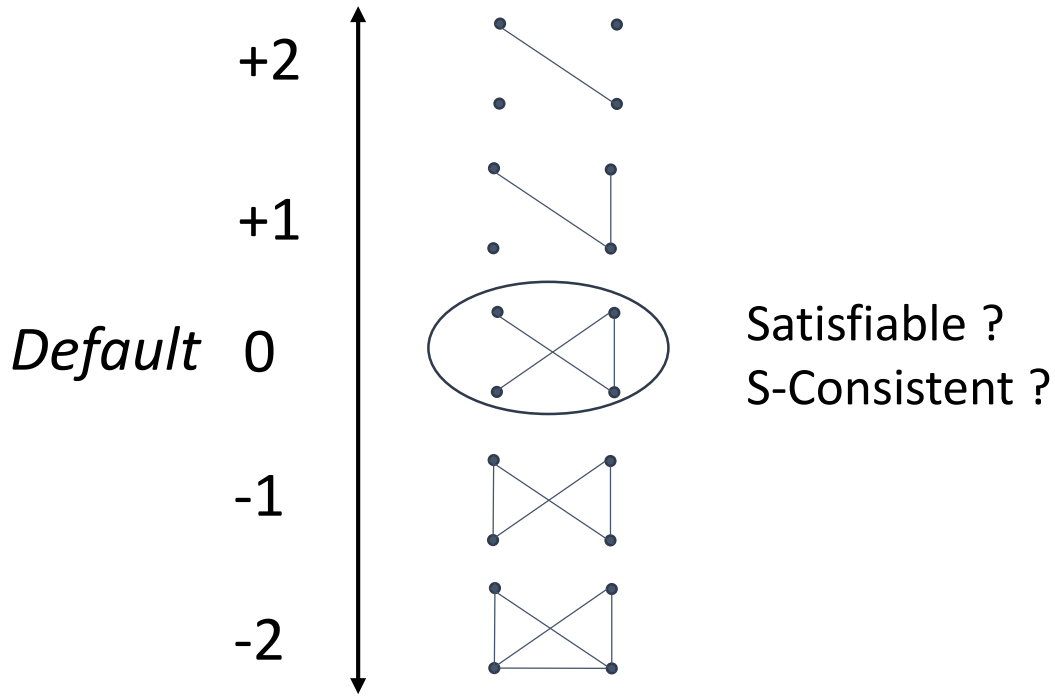
Stricter => Less orthologies



Looser => More orthologies

Experiments

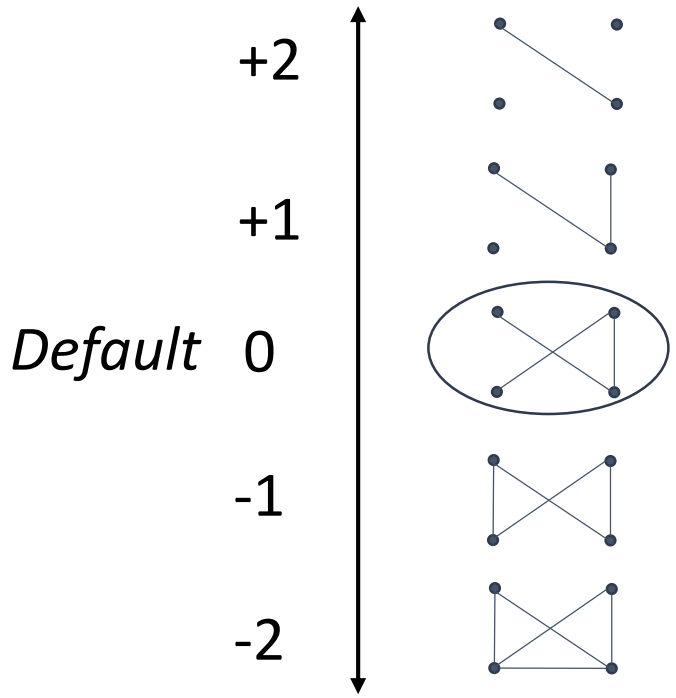
Stricter => Less orthologies



Looser => More orthologies

Experiments

Stricter => Less orthologies

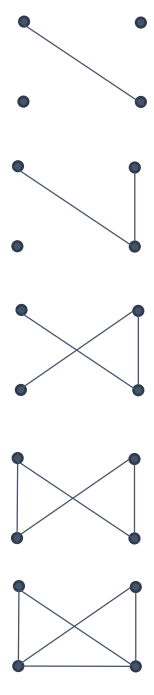


Satisfiable ? NO (~90% of families)
S-Consistent ? NO (~96% of families)

Looser => More orthologies

Experiments

Stricter => Less orthologies



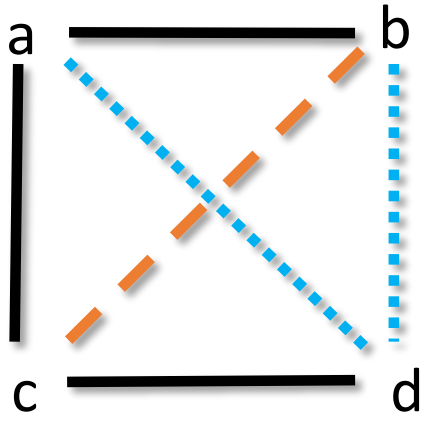
Looser => More orthologies

		NOT Satisfiable	NOT S-Consistent
	+2	80%	93%
	+1	82%	95%
Default	0	90%	96%
	-1	83%	95%
	-2	70%	89%

Unknown/undecided relations

We might lack confidence in some given relations

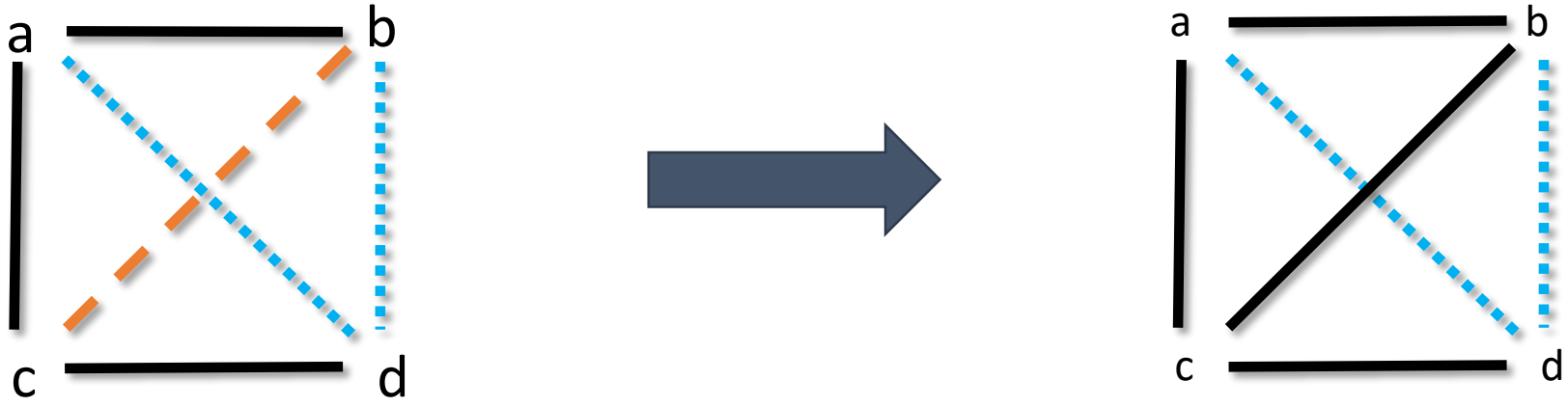
e.g. genes having a borderline BLAST similarity value



Problem :

Given a relation graph R with **unknown edges**, can they be **chosen** to make R:

- **satisfiable**?
- **S-Consistent**?
- **self-consistent**?



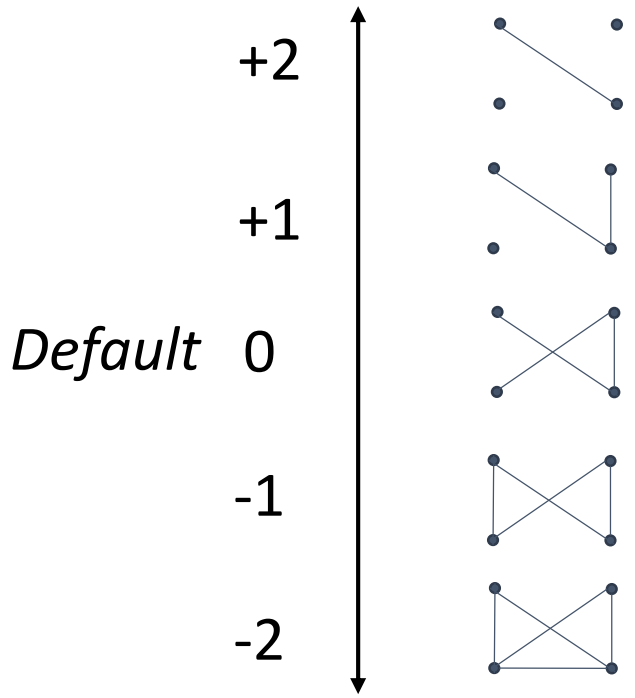
Problem :

Given a relation graph R with **unknown edges**, can they be **chosen** to make R:

- **satisfiable**? Polytime (Lafond & El-Mabrouk, 2014)
- **S-Consistent**? Polytime (Lafond & El-Mabrouk, 2014)

Experiments with the unknown

Stricter => Less orthologies

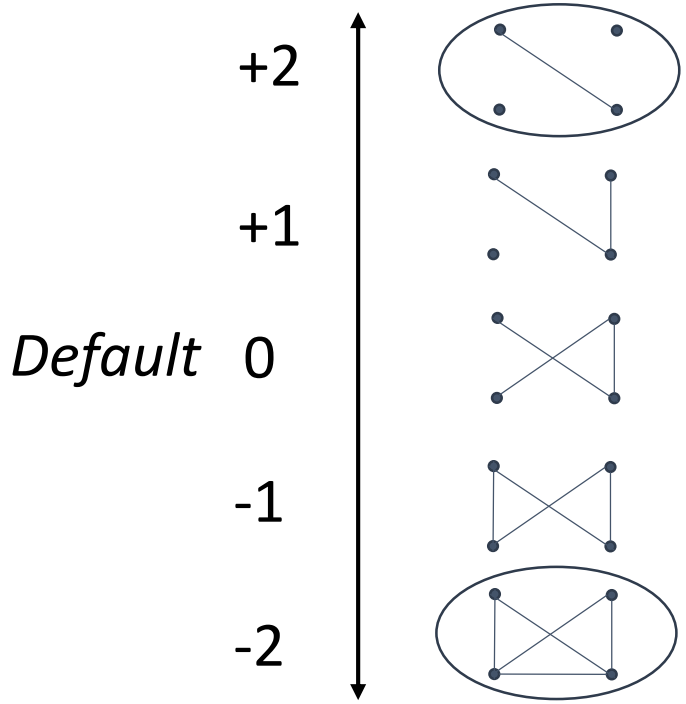


Looser => More orthologies

Can we get some robust relationships out of these ?

Experiments with the unknown

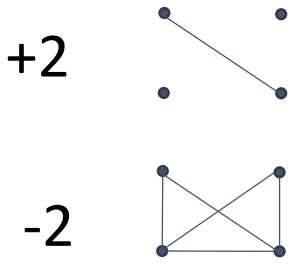
Stricter => Less orthologies



Looser => More orthologies

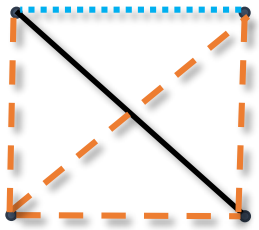
Can we get some robust relationships out of these ?

Experiments with the unknown

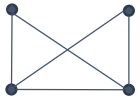
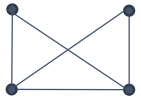


Keep the common orthologies and paralogies.

The rest is unknown.



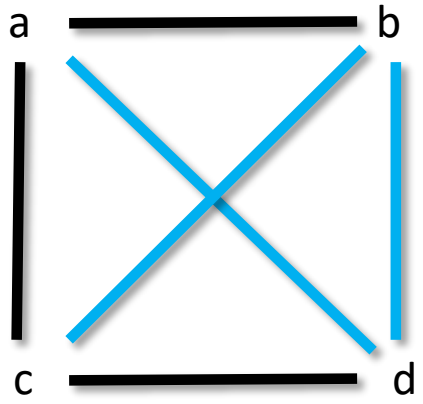
Experiments with the unknown

		NOT Satisfiable	NOT S-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%

Gene relation correction

Make R satisfiable by changing a minimum number of relations.

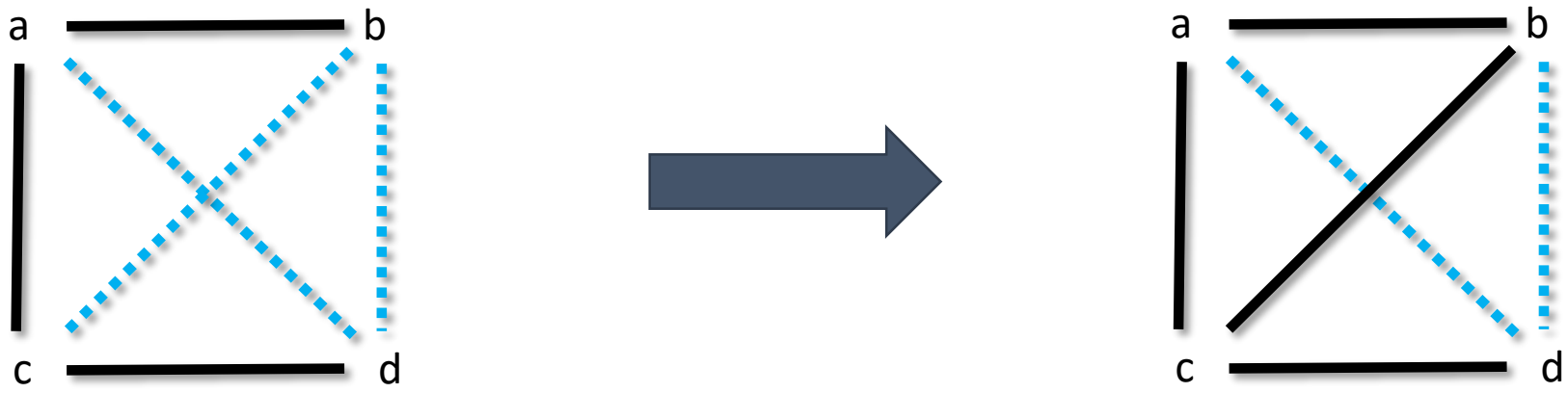
That is, change as few edge colors as possible to make R_{BLACK} P_4-free



Gene relation correction

Make R satisfiable by changing a minimum number of relations.

That is, change as few edge colors as possible to make R_{BLACK} P_4-free

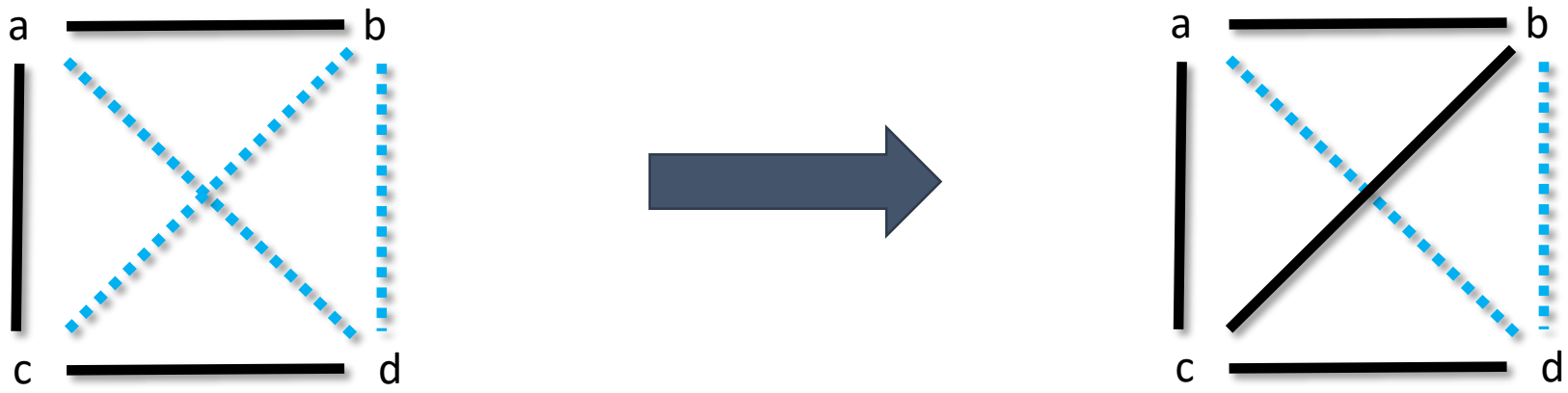


Gene relation correction

Make R satisfiable by changing a minimum number of relations.

That is, change as few edge colors as possible to make R_{BLACK} P_4-free

NP-Complete (El-Mallah & Colbourn, 1988)



Gene relation correction

- Many other variants, all difficult:
  - Remove as few genes to have a P4-free graph => can't even approximate
  - Incorporate information from species tree => still NP-complete
  - Add weights on the orthology/paralogy relations => can't approximate

  (Dondi, Lafond, El-Mabrouk, 2014-2016)

ILP formulation (has difficulty handing > 10 genes)

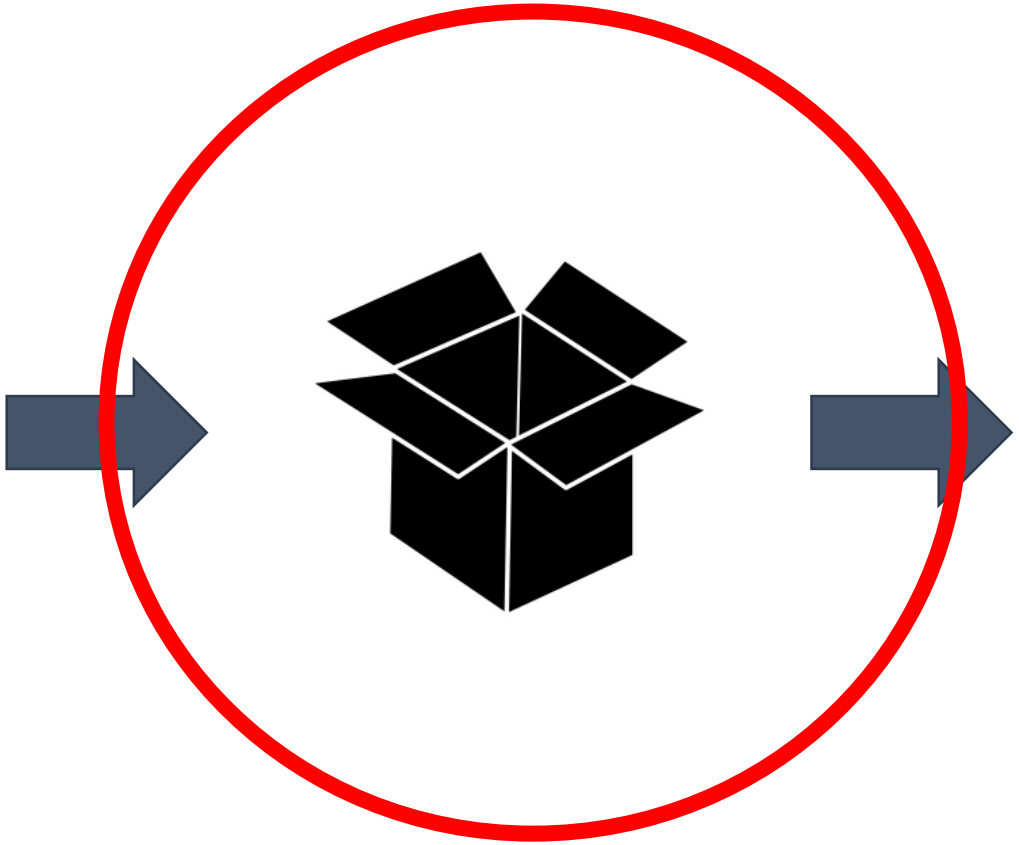
FPT algorithms (also slow)

MinCut heuristic (no performance guarantees)

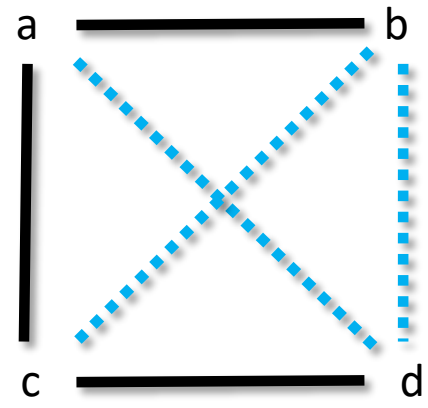
Dealing with similarity-based methods

Orthology/paralogy relation graph R

Sequences and stuff



R

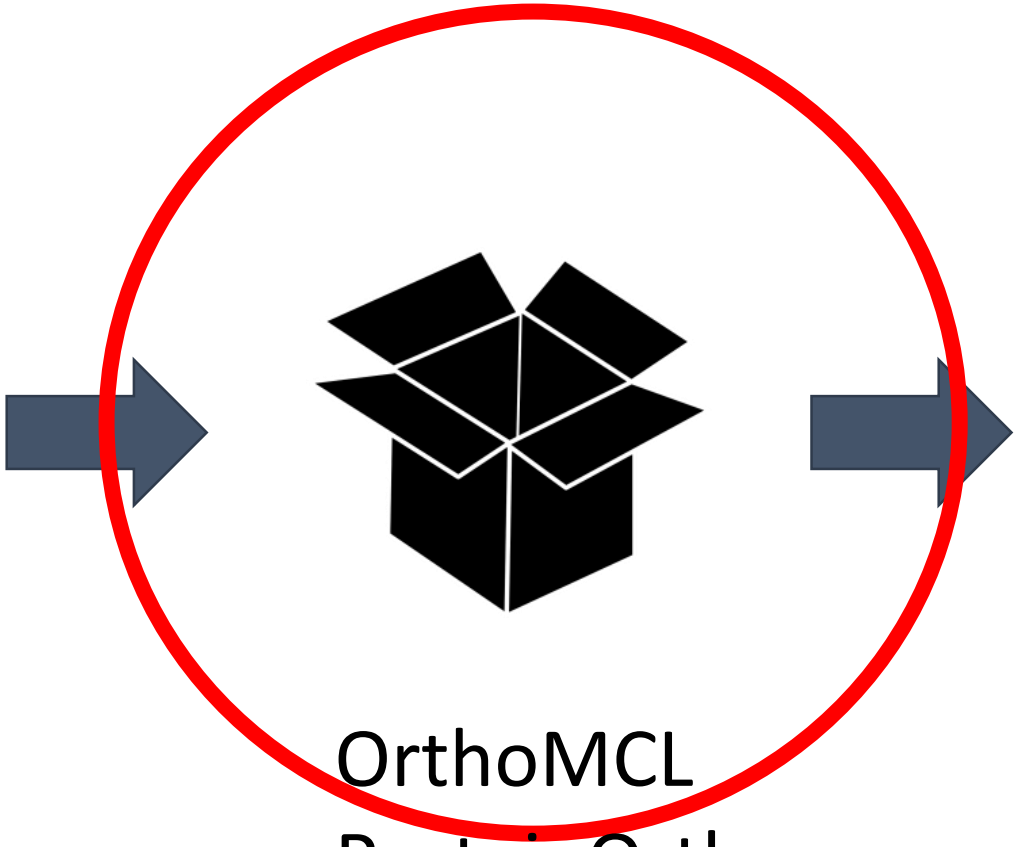


Orthologs = (a,b) (a, c) (c, d)

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

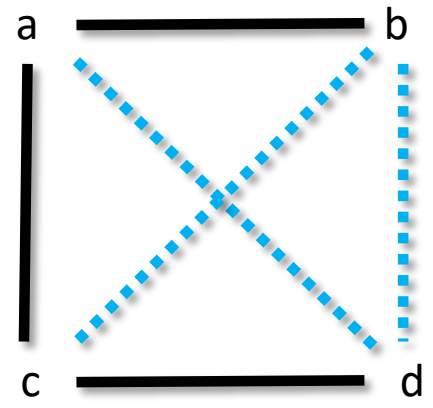
Orthology/paralogy relation graph R

Sequences and stuff



OrthoMCL
ProteinOrtho
OrthoFinder
...

R



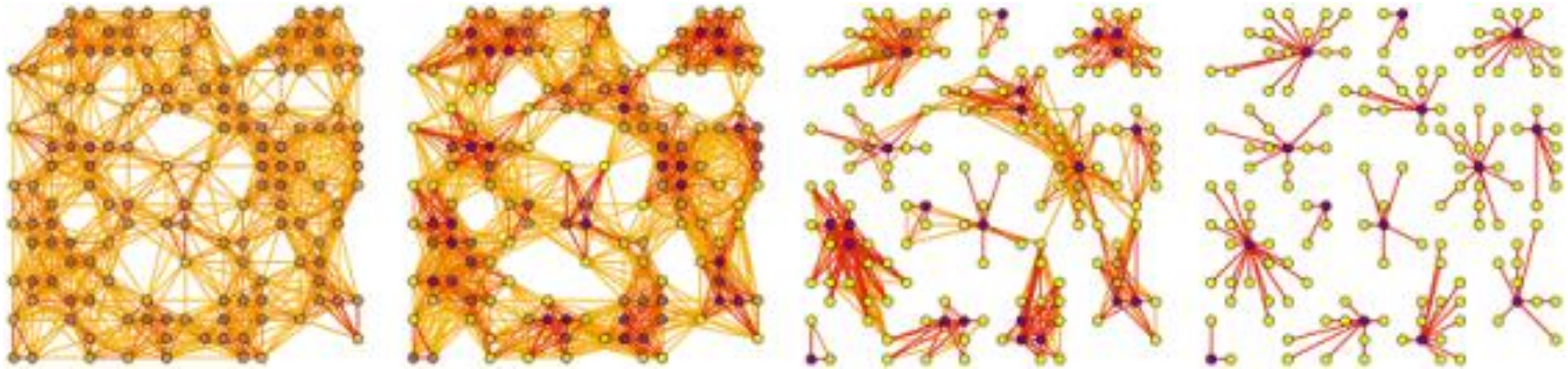
Orthologs = (a,b) (a, c) (c, d)

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

Traditional inference method

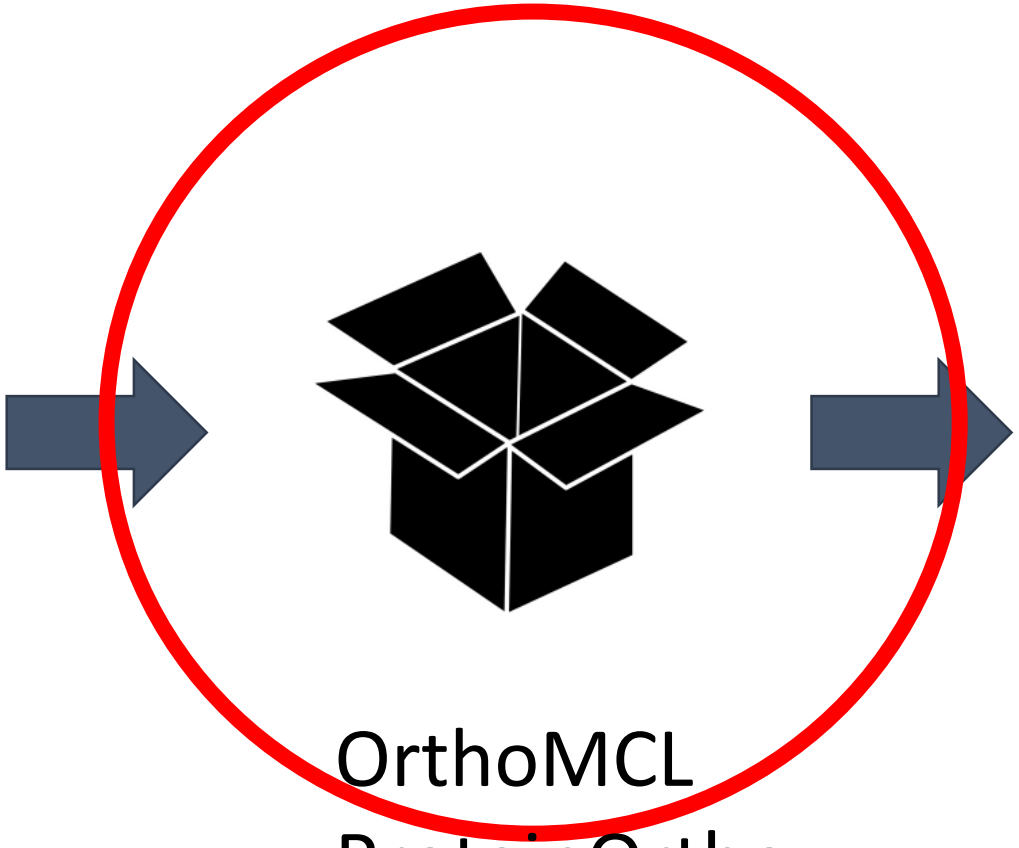
Clustering genes into groups of orthologs:

- If g1 and g2 and "**similar enough**" in terms of sequence, we say that g1 and g2 are putative orthologs
- "Similar enough" usually means that, if g1 and g2 are from species s1 and s2, they for a Bidirectional Best Hit (BBH):
  - g1's best match in s2 is g2
  - g2's best match in s1 is g1

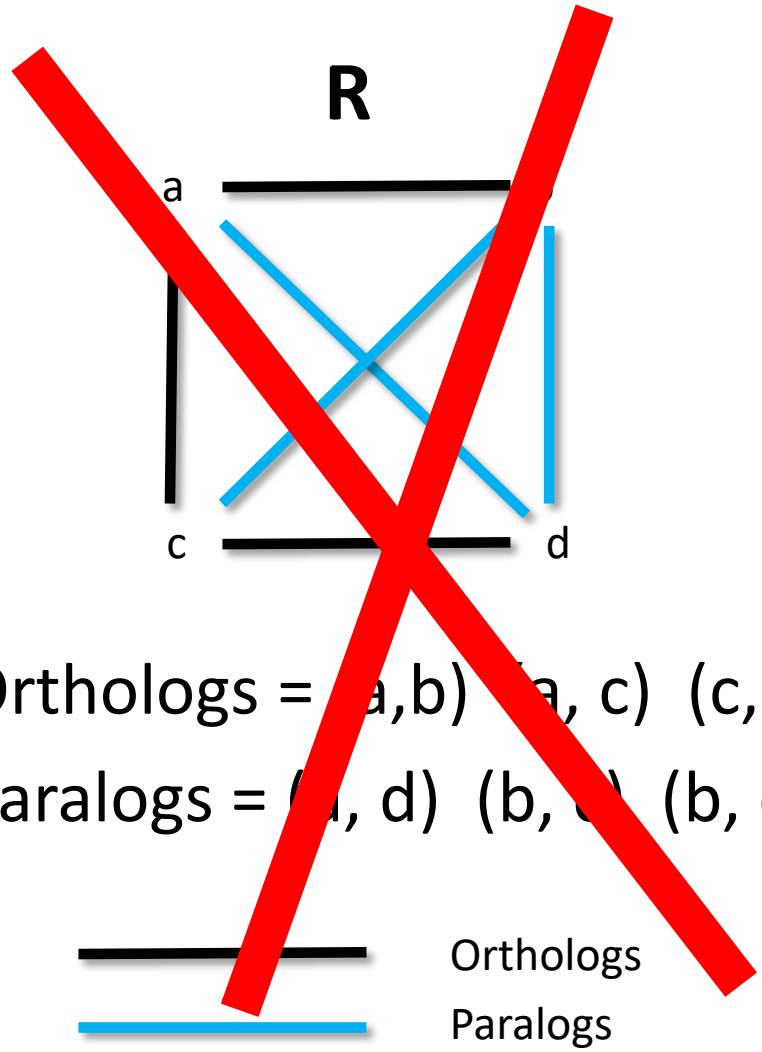


Orthology/paralogy relation graph R

Sequences and stuff



OrthoMCL
ProteinOrtho
OrthoFinder
…

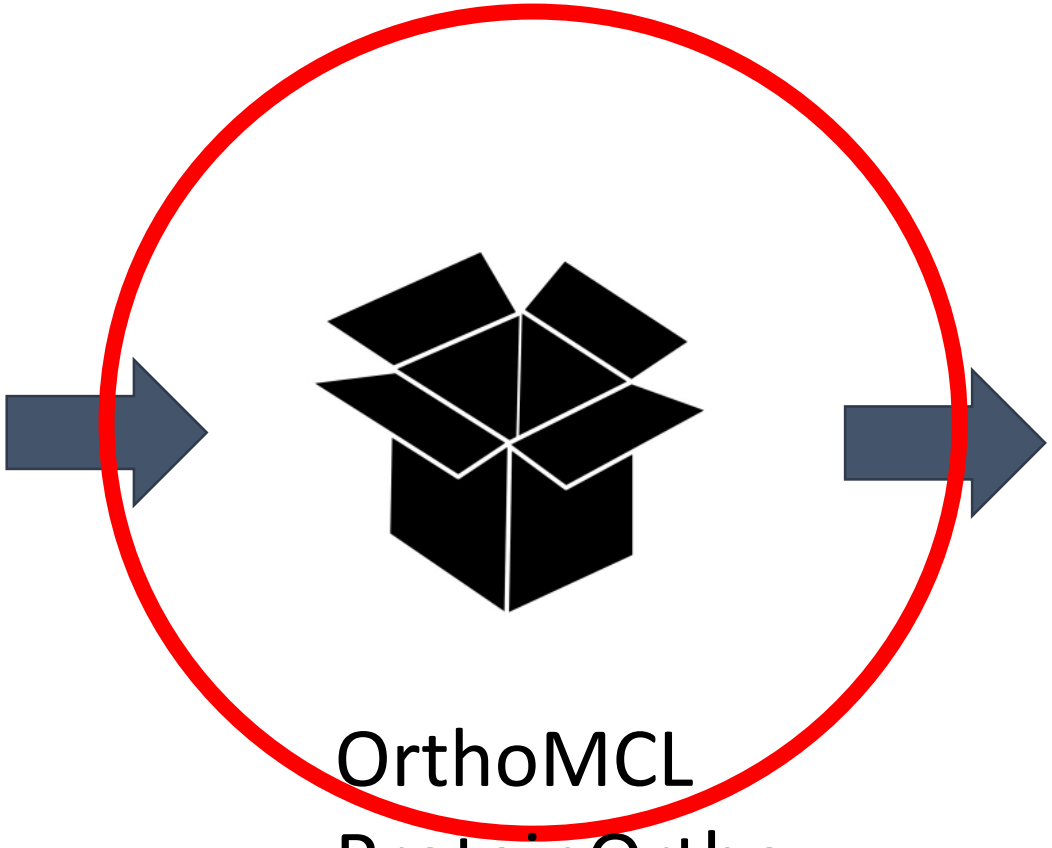


Orthologs = (a,b) (a, c) (c, d)

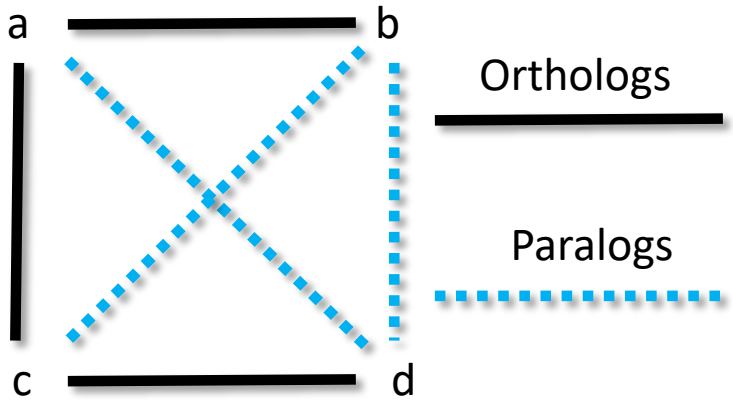
Paralogs = (a, d) (b, c) (b, d)

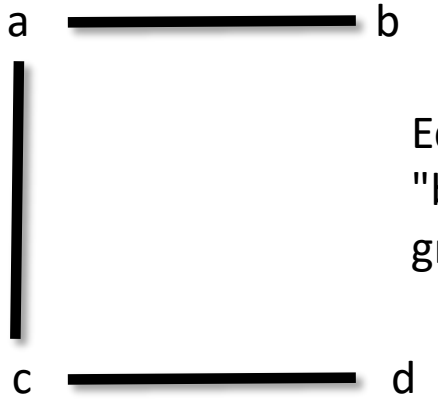
Relation graph vs similarity graph

Sequences and stuff



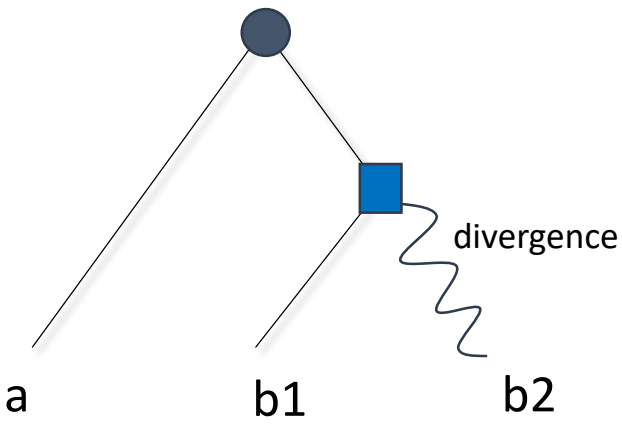
OrthoMCL
ProteinOrtho
OrthoFinder
…

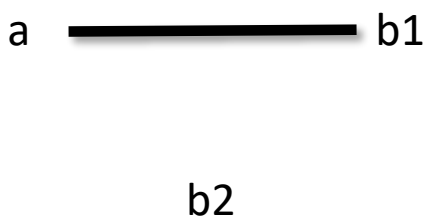




Edge = "similar", or "belong ot the same group"

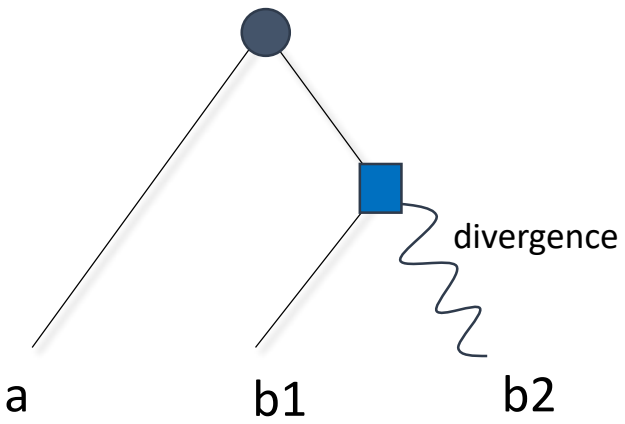
Dup after speciation is confusing

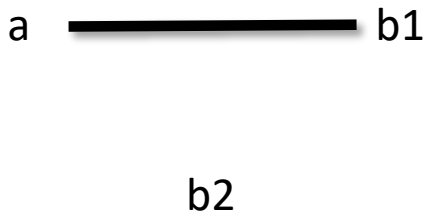




Similarity graph

Dup after speciation is confusing

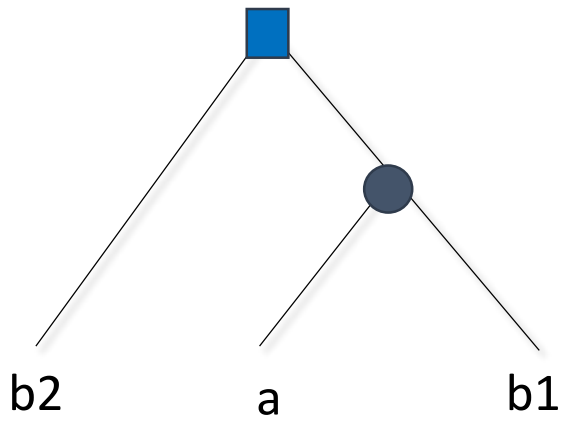




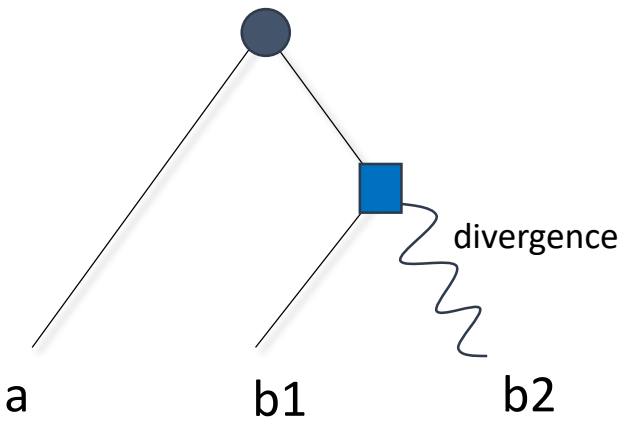
Similarity graph

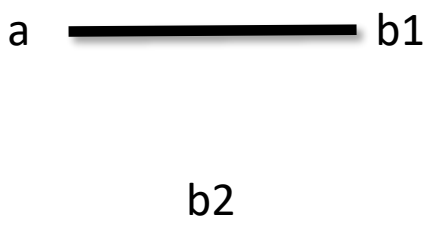
Interpreted as a relation graph:
(a, b1) = orthologs
(a, b2) = paralogs
(b1, b2) = paralogs

Gene tree for these relations



Dup after speciation is confusing

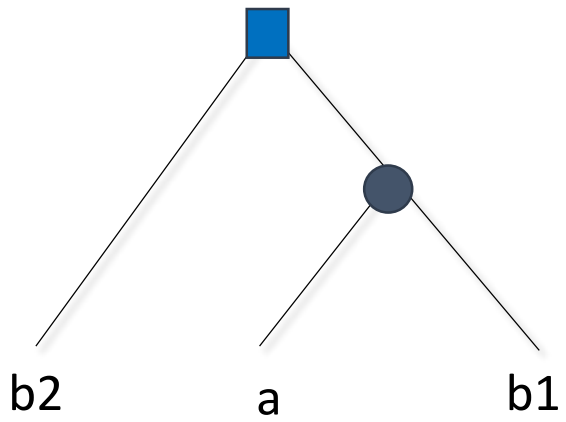




Similarity graph

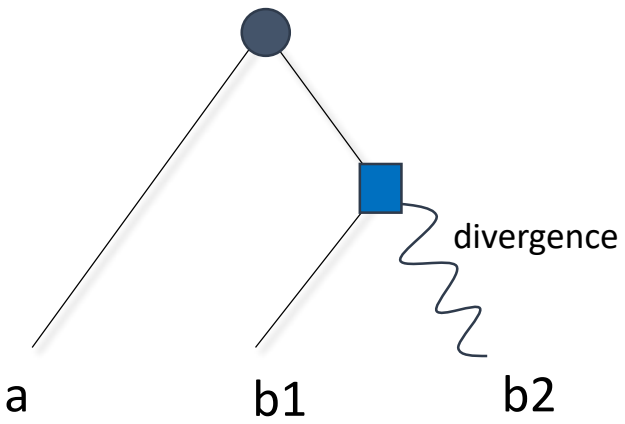
Interpreted as a relation graph:
(a, b1) = orthologs
(a, b2) = paralogs
(b1, b2) = paralogs

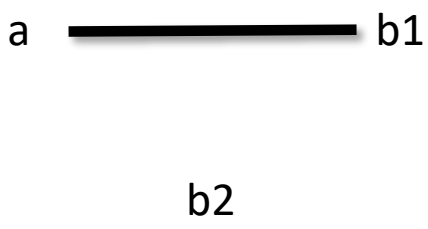
Gene tree for these relations



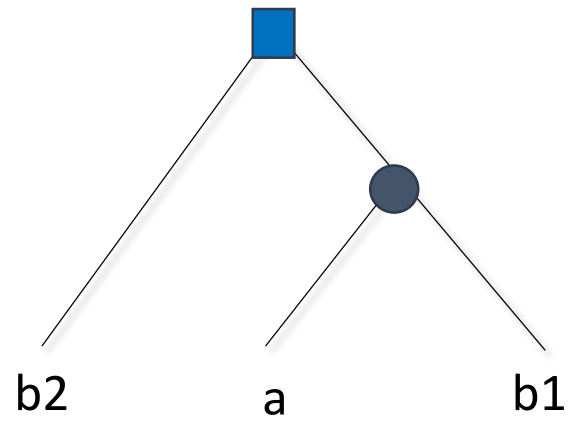
The (a, b2) orthology is indistinguishable from paralogy from the point of view of similarity.

Dup after speciation is confusing





Interpreted as a relation graph:
(a, b1) = orthologs
(a, b2) = paralogs
(b1, b2) = paralogs



BAD for:

1) **Benchmarking: the graph passes the test of being P4-free, and yet does not depict relations correctly**
2) **Gene tree reconstruction: interpreting as relations yields the wrong gene tree.**

Some options to address this issue

1) **Give up on these missing orthologs.**
2) **Devise methods that really infer relation graphs.**
3) **Deal with the similarity graphs.**

Some options to address this issue

1) **Give up on these missing orthologs.**
2) **Devise methods that really infer relation graphs.**
3) **Deal with the similarity graphs.**

Some options to address this issue

1) **Give up on these missing orthologs.**
2) **Devise methods that really infer relation graphs.**
3) **Deal with the similarity graphs.**
   - **Can we characterize "valid" similarity graphs, analogously as what we did with relation graphs?**
   - **Yes, they are called leaf-powers by the graph theorists.**

Some options to address this issue

1) **Give up on these missing orthologs.**
2) **Devise methods that really infer relation graphs.**
3) **Deal with the similarity graphs.**
   - **Can we characterize "valid" similarity graphs, analogously as what we did with relation graphs?**
   - **Yes, they are called leaf-powers by the graph theorists.**
     - **Recognizing leaf-powers is a longstanding open problem (not known to be in P nor NP-complete)**

Some options to address this issue

1) **Give up on these missing orthologs.**
2) **Devise methods that really infer relation graphs.**
3) **Deal with the similarity graphs.**
   - **Can we characterize "valid" similarity graphs, analogously as what we did with relation graphs?**
   - **Yes, they are called leaf-powers by the graph theorists.**
     - **Recognizing leaf-powers is a longstanding open problem (not known to be in P nor NP-complete)**
   - **Too complicated, let's start with a restricted model**

The Divergence-After-Duplication (DAD) model

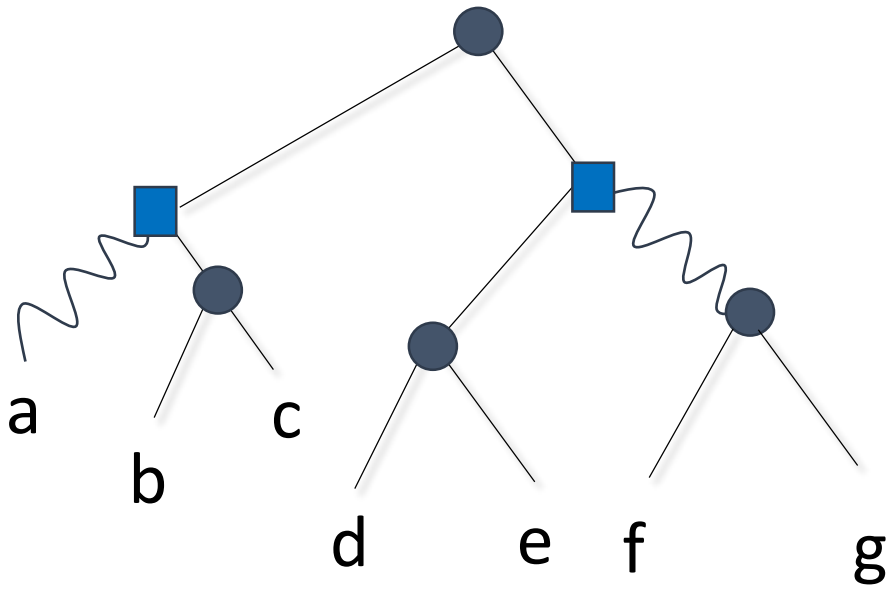
Orthologs conjecture: orthologous genes tend to be similar in function, whereas paralogous genes tend to differ.

The Divergence-After-Duplication (DAD) model

1) In the absence of gene duplication, no significant dissimilarity should be observed.
2) In the event of gene duplication, one copy remains intact whereas the other evolves at an accelerated rate.

(as in the motivation for the orthologs conjecture)

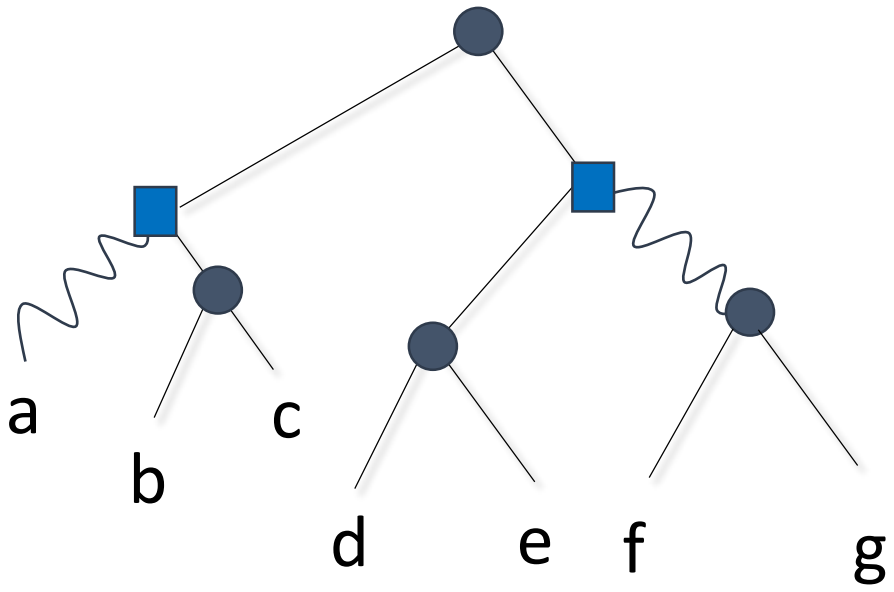
The Divergence-After-Duplication (DAD) model



Direct consequences of the axioms of the DAD model:

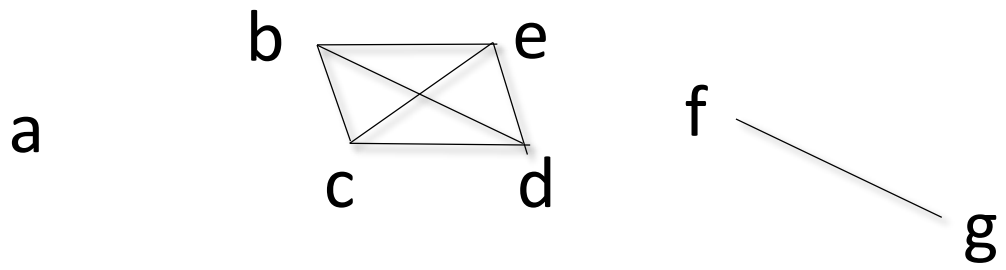
- Two genes will appear as "non-similar" **if and only if** a divergent duplication edge separates them.
- The similarity graph should contain **nothing else than cliques**.

The Divergence-After-Duplication (DAD) model

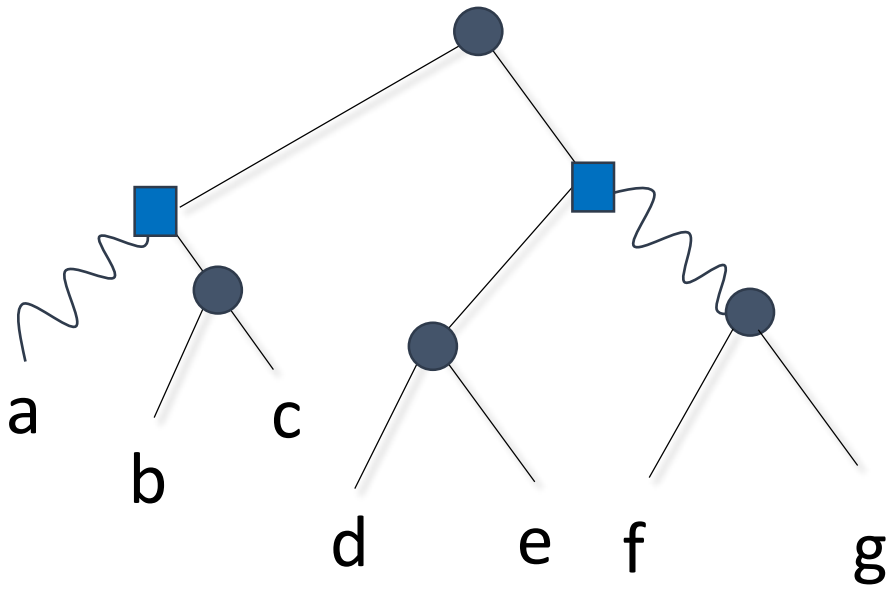


Direct consequences of the axioms of the DAD model:

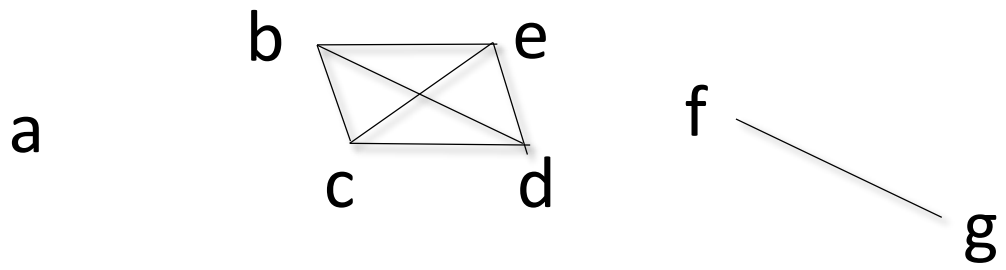
- Two genes will appear as "non-similar" **if and only if** a divergent duplication edge separates them.
- The similarity graph should contain **nothing else than cliques**.



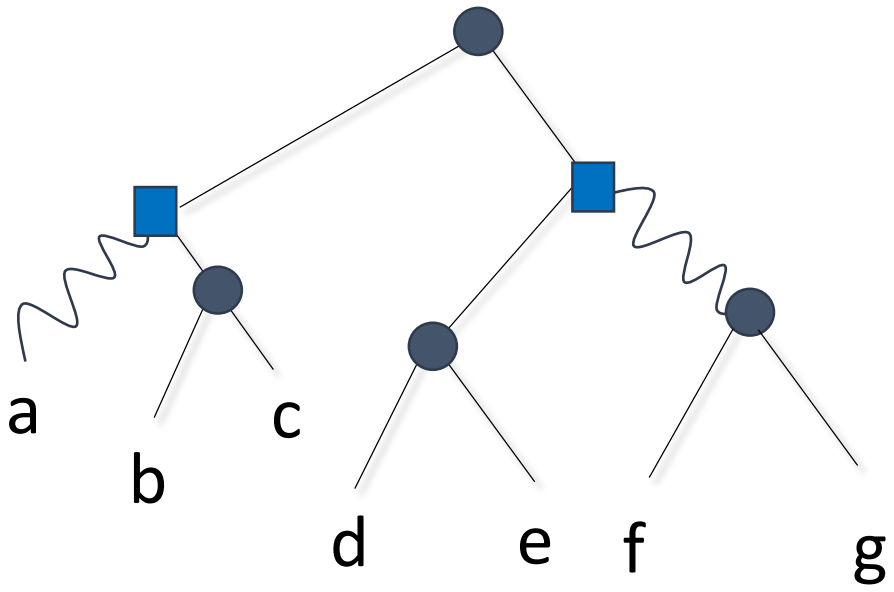
The Divergence-After-Duplication (DAD) model



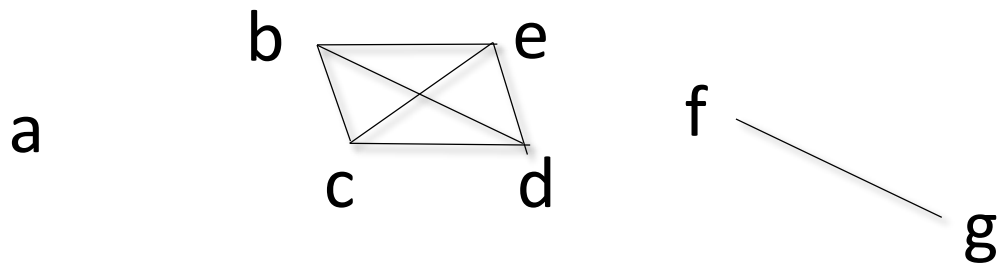
- Clustering algorithms can be applied to find the **"similarity cliques"**, which we assume represent orthology subtrees.
- The cliques do not represent all orthologies: some (and perhaps many) **may be missing**, e.g. (b, f), (b, g), (c, f), …



The Divergence-After-Duplication (DAD) model



- Clustering algorithms can be applied to find the **"similarity cliques"**, which we assume represent orthology subtrees.
- The cliques do not represent all orthologies: some (and perhaps many) **may be missing**, e.g. (b, f), (b, g), (c, f), …
- **How can we find missing relations?**
  - **(WIP)**



Conclusion

- **Orthology/paralogy** graphs are exactly the P_4-free graphs
- In practice, we only have a **similarity graph**
  - Not the same
- Can we "turn" a similarity graph into an orthology/paralogy graph?
  - What are the limits of similarity for orthology inference?
- Future works: design algorithms to infer missing orthologs from a similarity graph, and test them on real/simulated datasets.

Gene relation correction

Make R **S-Consistent** by changing a minimum number of relations.

That is, change as few edges colors so that R is P_4-free, and every P_3 agrees with S. (hey, maybe S can help reduce the complexity)

Gene relation correction

Make R **S-Consistent** by changing a minimum number of relations.

That is, change as few edges colors so that R is P_4-free, and every P_3 agrees with S. (hey, maybe S can help reduce the complexity)

NO

NP-Complete (Lafond & El-Mabrouk, 2014)

Gene relation correction

Make R **S-Consistent** by removing a minimum **number of genes**.

That is, **delete as few vertices** from R so that R is P_4-free, and every P_3 agrees with S.

Gene relation correction

Make R **S-Consistent** by removing a minimum **number of genes**.

That is, **delete as few vertices** from R so that R is P_4-free, and every P_3 agrees with S.

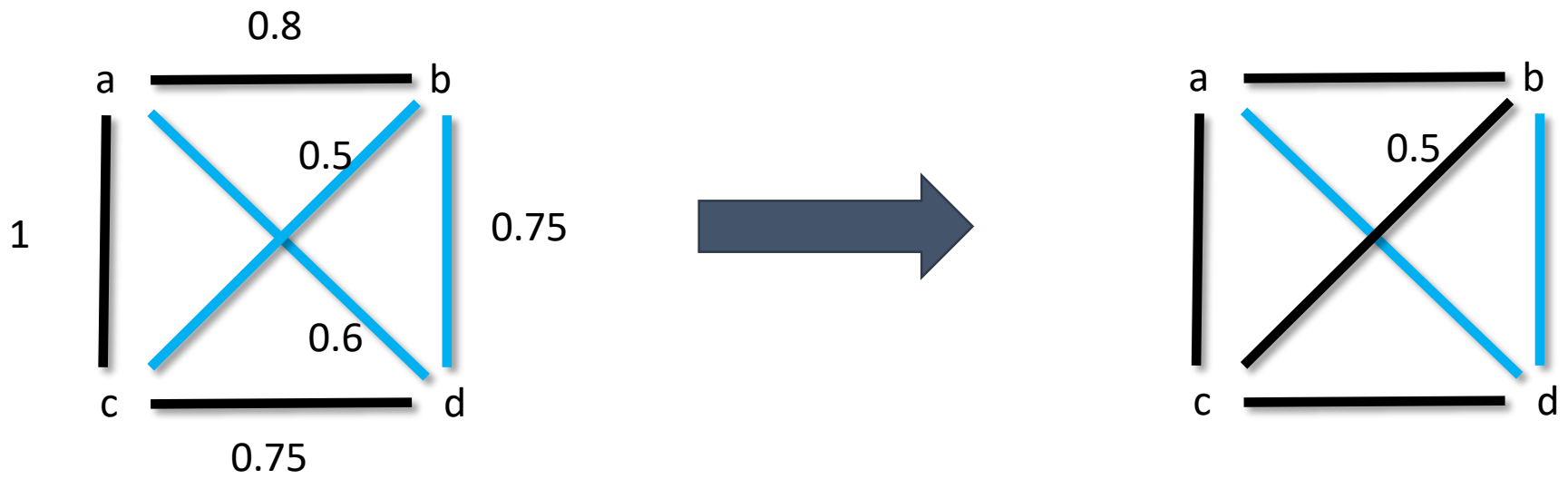
NP-Hard to approximate within a $n^{1-\varepsilon}$ factor. (Lafond, Dondi, & El-Mabrouk, 2016)

Weighted gene relation correction

To make things easier:

Give each edge a **weight**, representing some degree of confidence over the inferred orthology/paralogy.

This weight represents the cost for changing the edge's color.



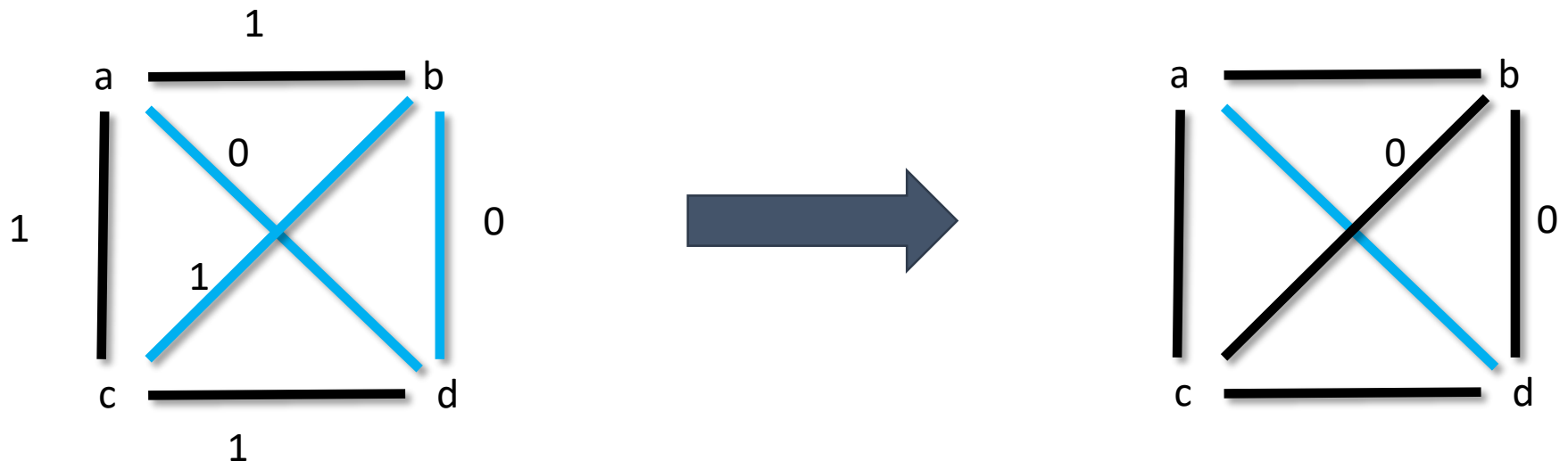
Weighted gene relation correction

Something we can handle:

If edges all have weights of 0 or 1

0 = don't care, 1 = don't touch

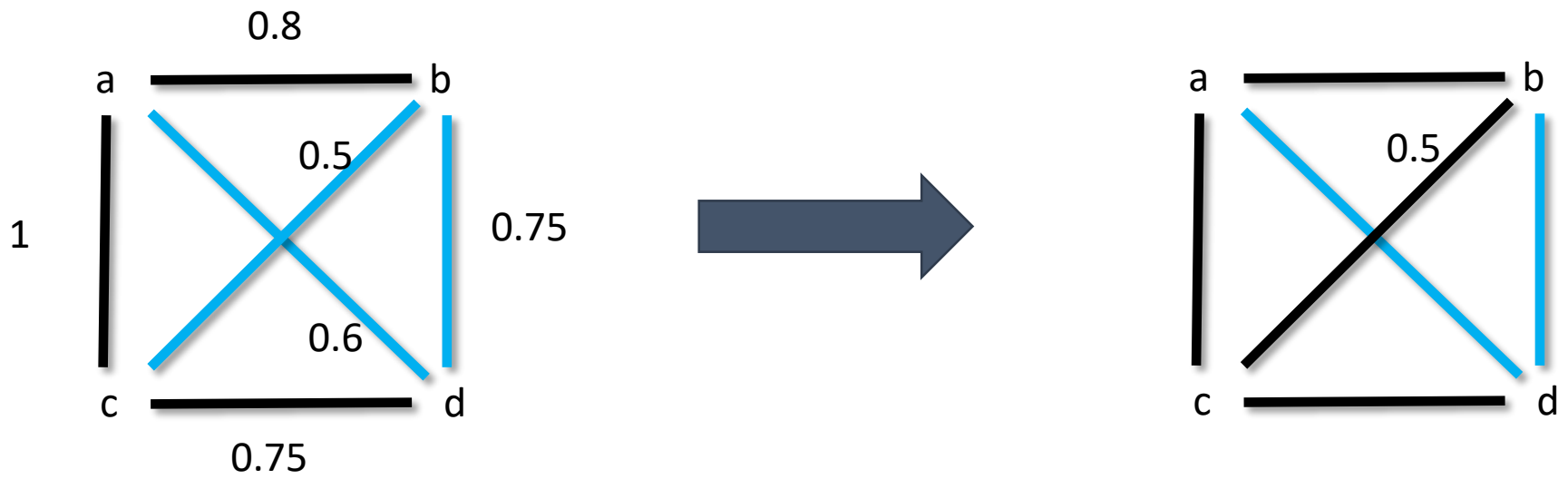
We can tell in polynomial time if there is an edge editing of weight 0.



Weighted gene relation correction

If weights are arbitrary, NP-Hardness follows from the unweighted version (for both satisfiability and consistency).

Worse than that, there is **no constant factor approximation** assuming the *unique games conjecture*.



Fixed parameter tractability

k = number of edges that can be edited

For **satisfiability**, the unweighted edge-editing problem admits a vertex kernel of size $O(k^3)$ (Guillemot, Paul, Perez, 2010)

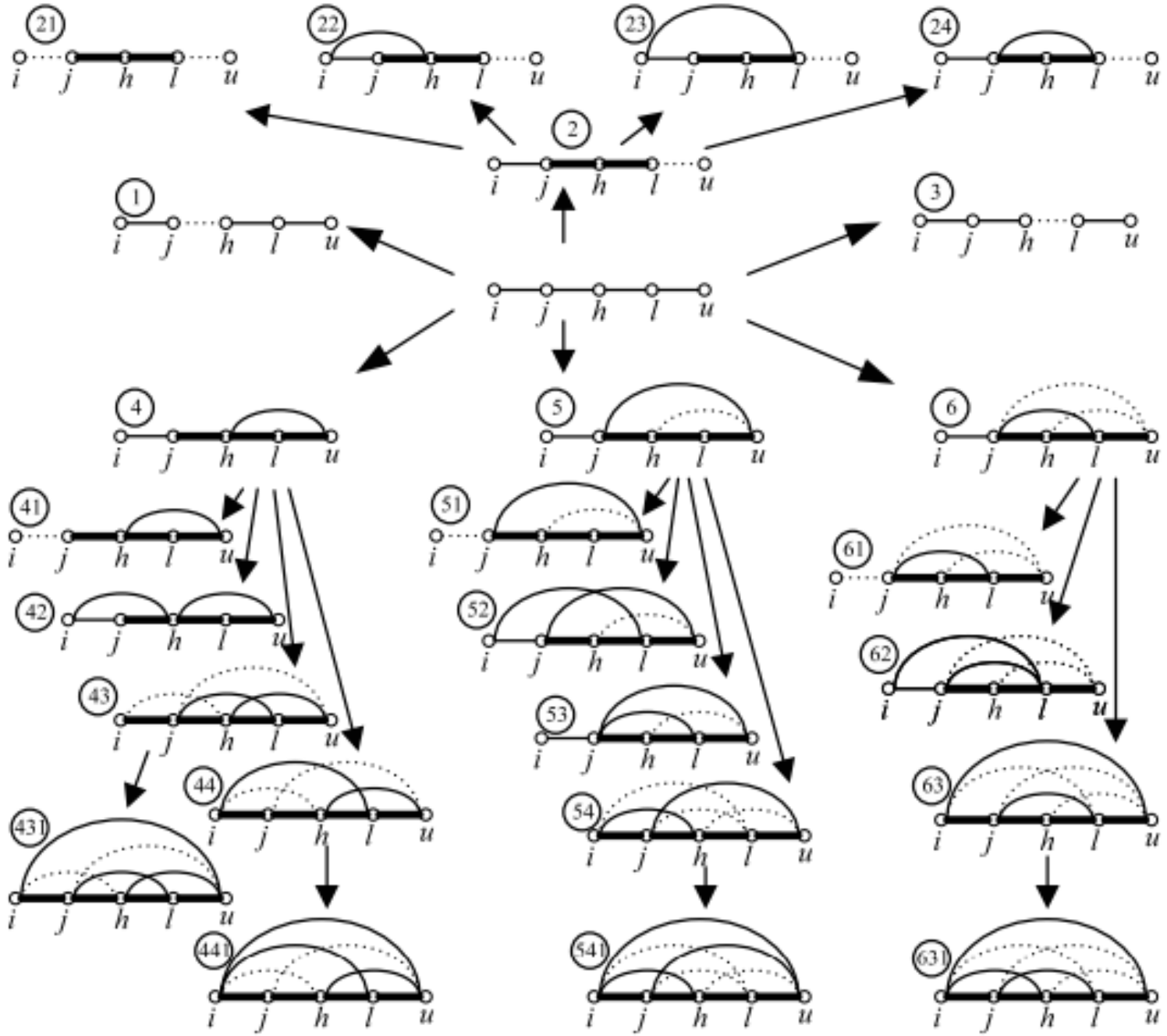
There is an obvious FPT algorithm:

each P_4 must be killed. There are 6 edge modifications that accomplish this. Branch into each possibility. $O(6^k n)$

- can be extended to S-consistency

Was improved to $O(4.612^k + |V|^{4.5})$ (Lui, Wang, Guo & Chen, 2012)

- good for S-consistency? No idea.


21
22
23
24
2
1
3
4
5
6
41
51
61
42
52
62
43
53
63
431
44
54
441
541
631
i j h l u

Min-cut approximation for satisfiability

Recall:

Theorem (again):

A relation graph R is satisfiable if and only if for each subgraph R', one of R'_{BLACK} or R'_{BLUE} is disconnected.

In particular, R_{BLACK} or its complement R_{BLUE} must be disconnected.

So we'll disconnect it then.

Min-cut approximation for satisfiability

In particular, R_{BLACK} or its complement R_{BLUE} must be disconnected.

Find a min-cut on R_{BLACK}

Find a min-cut on R_{BLUE}

Take the best of the two and apply.

Repeat on the resulting components.

(min-cut = minimum weight edge-set that disconnect R, can be found in time $O(n^3)$)

Min-cut approximation for satisfiability

In particular, R_{BLACK} or its complement R_{BLUE} must be disconnected.

Find a min-cut on R_{BLACK}

Find a min-cut on R_{BLUE}

Take the best of the two and apply.

Repeat on the resulting components.

(min-cut = minimum weight edge-set that disconnect R, can be found in time $O(n^3)$)

Gives a solution that is at most n times worse than optimal.

(not great, but shows that approximability is bounded)

Theoretical and practical problems

Theoretical problems

Unweighted case: can we approximate satisfiability? Consistency?

Weighted case: gap in approximability results. Is there better than a n-factor approximation? Somewhere in-between constant and n.

FPT : elements of unweighted satisfiability correction (aka cograph-editing) are known. Not much about the rest.

Practical problems

How do we even infer **orthology and paralogy**?

(but earlier I said we could!)

However, similarity-based approaches form clusters of orthologs.

Not exactly the same thing.

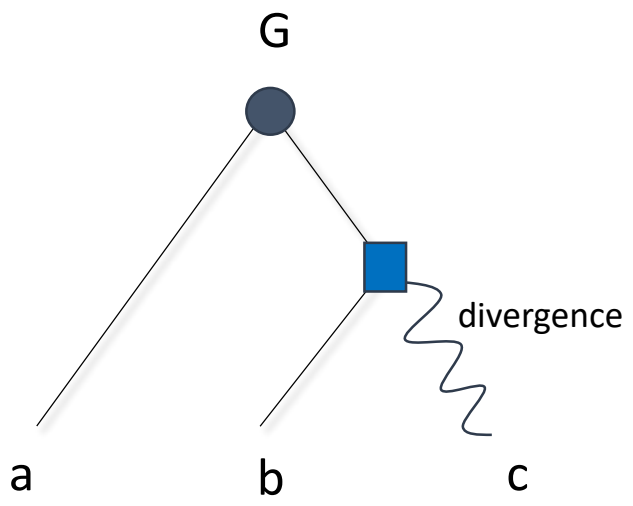
Practical problems

How do we even infer **orthology and paralogy**?

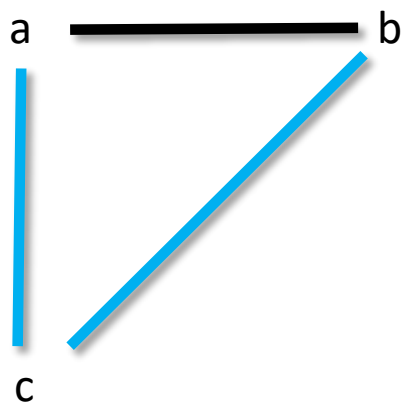
(but earlier I said we could!)

However, similarity-based approaches form clusters of orthologs.

Not exactly the same thing.



Similarity graph =/= orthology/paralogy graph



Practical problems

We don't even know **how to test** our correction methods.

- Gold standard datasets are extremely rare, if nonexistent.
- Most software are interested into forming clusters of orthologs. How do we compare with others?

Practical problems

Faster approximations and heuristics are still needed.

The Min-Cut algorithm takes time $O(n^3)$, and our implementation is too slow for, say, 1000 genes.

How to handle **other events**?

How can we distinguish species tree disagreement with HGT or ILS? Beyond graph theory, what is their practical impact in the ortholgoy/paralogy inference process?