Algorithms for the validation and correction of gene relations

Manuel Lafond, Université de Montréal

Introduction

Gene trees, species trees

Duplication, speciation

Orthologs, paralogs, and why?

Validation of relations

Cograph (P_4-free) characterization of valid relations

Relations consistent with a species tree

Relation correction

Open theoretical and practical problems

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.

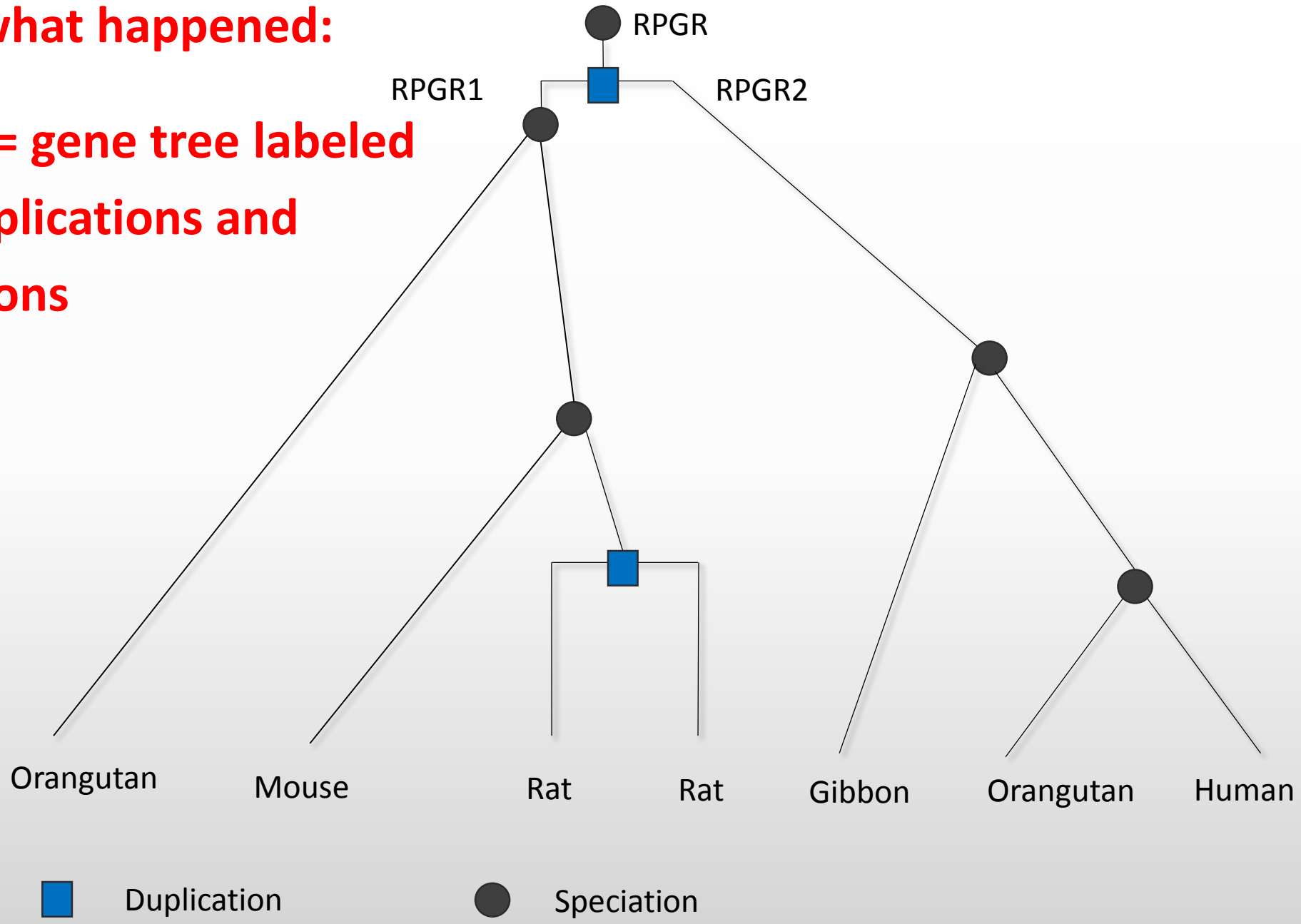
What happened exactly?

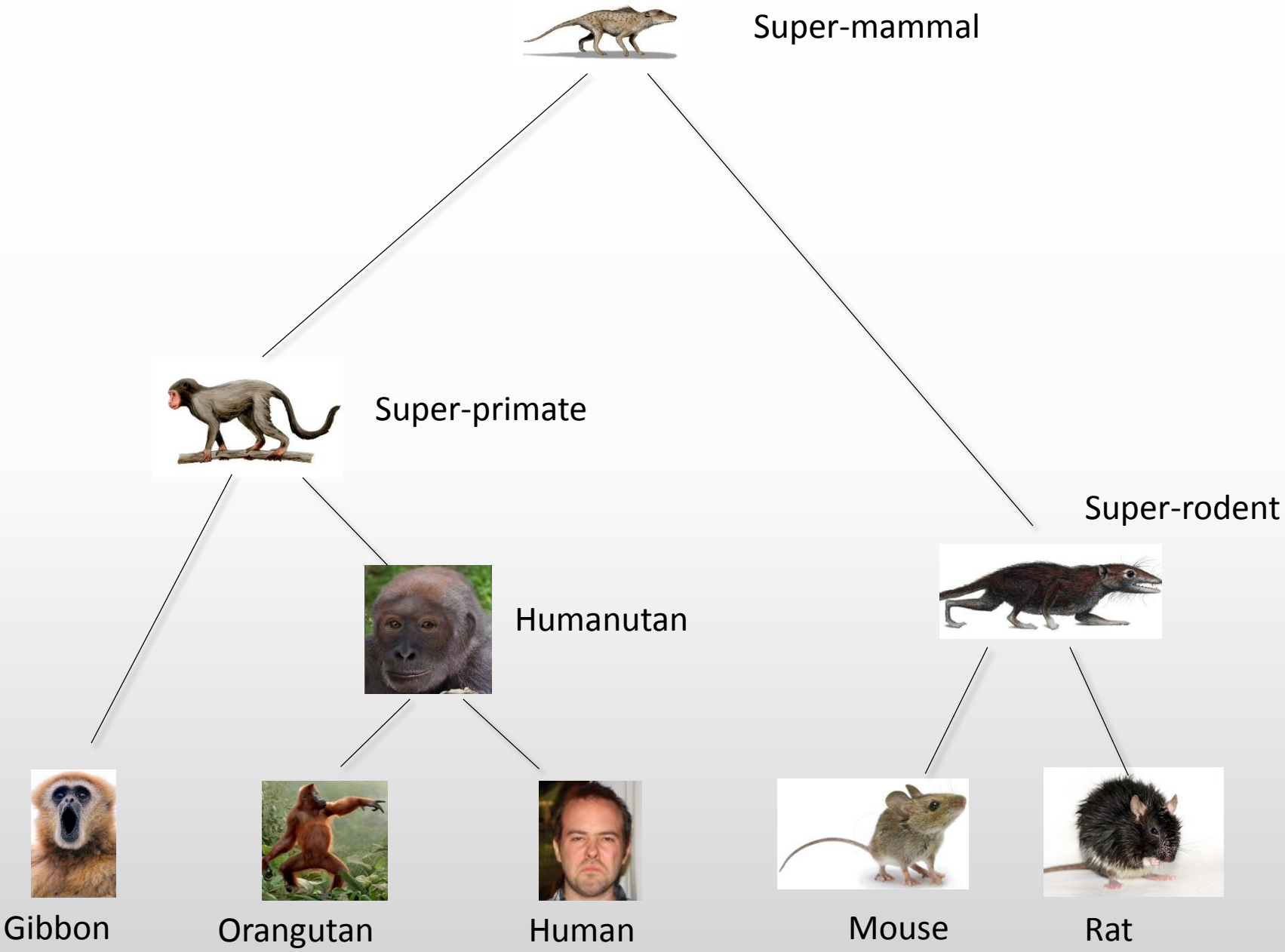
A gene can be :

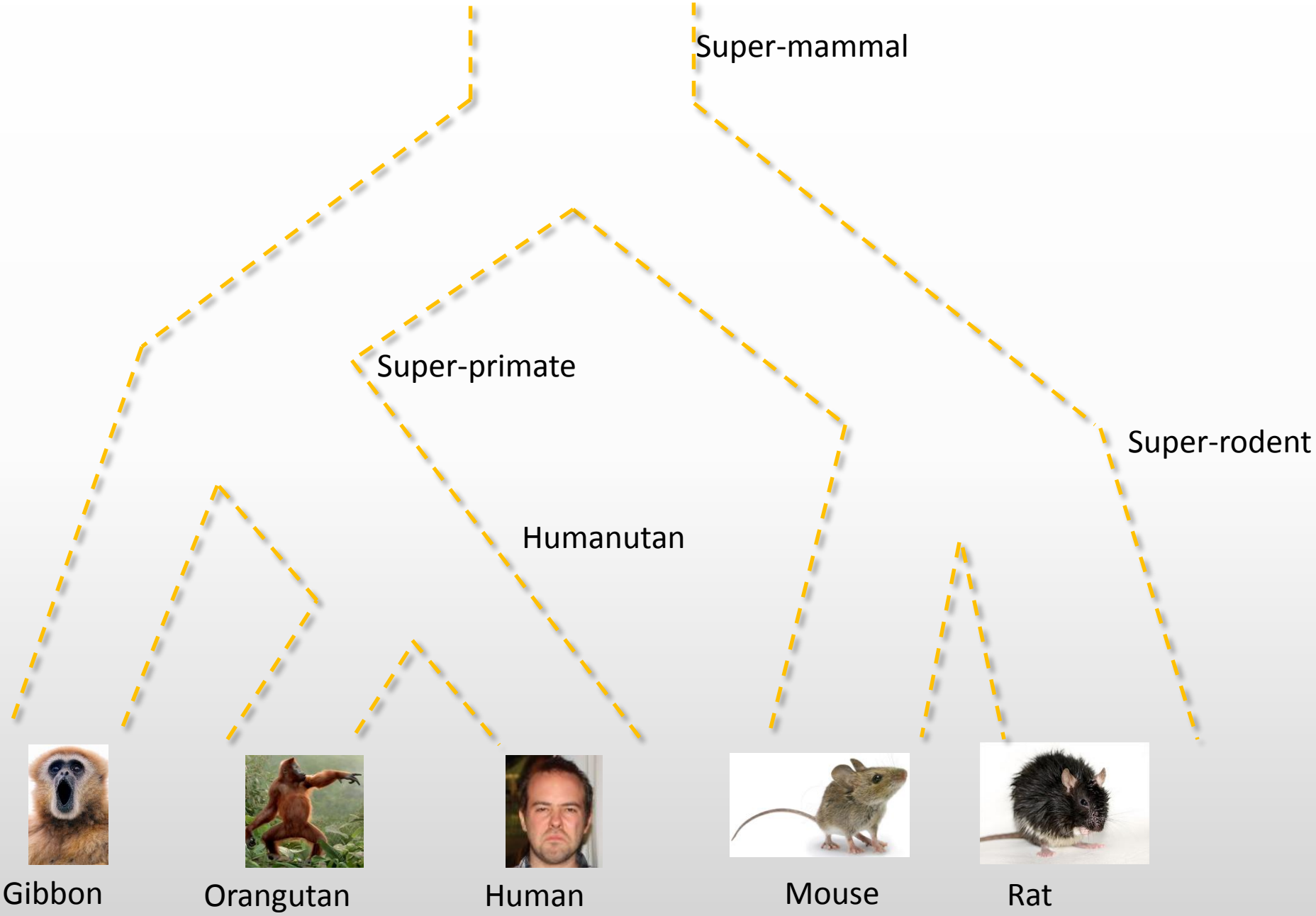
- **Transmitted** to descending species by **speciation**
- **Duplicated**
- **Lost**

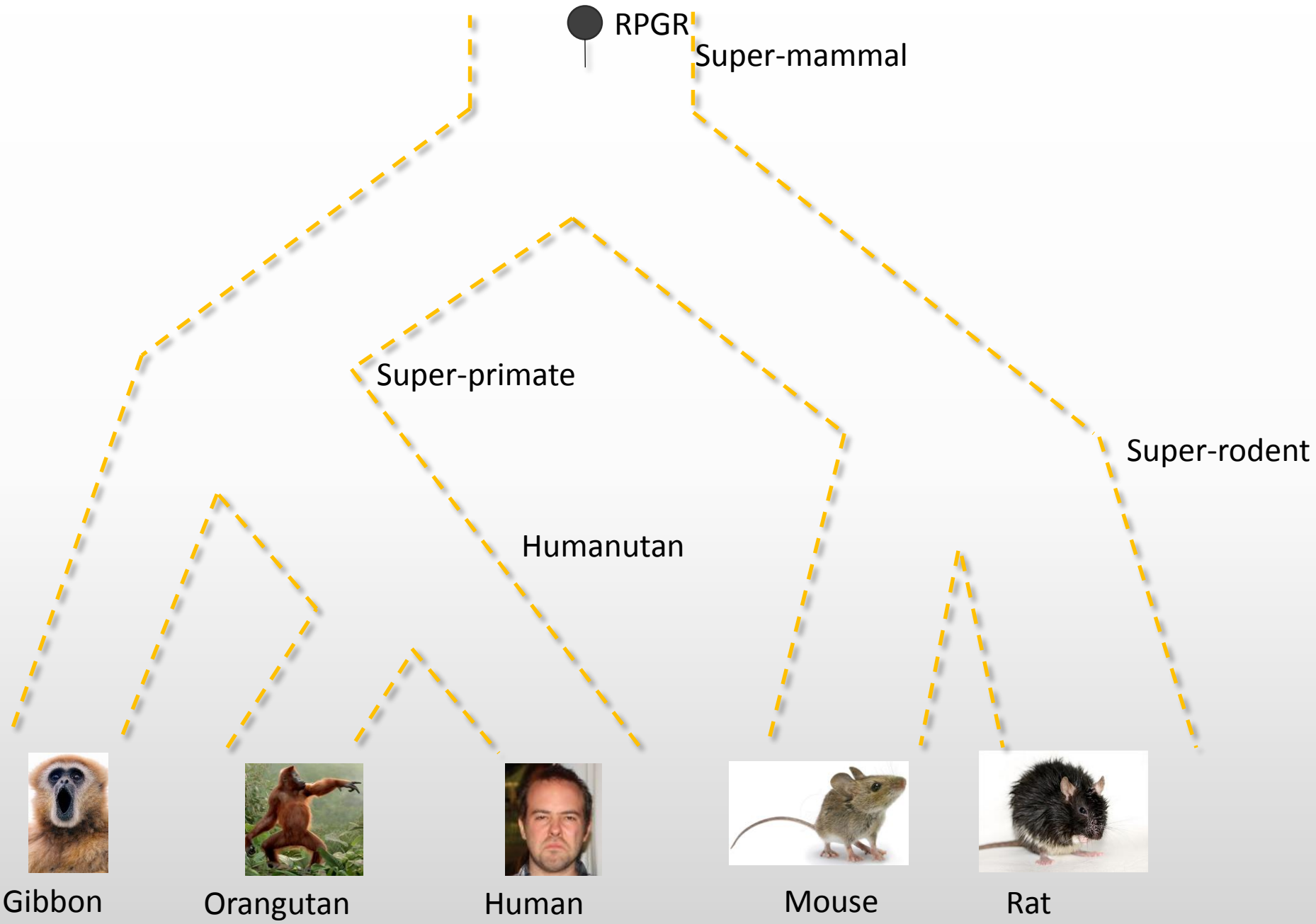
Here's what happened:

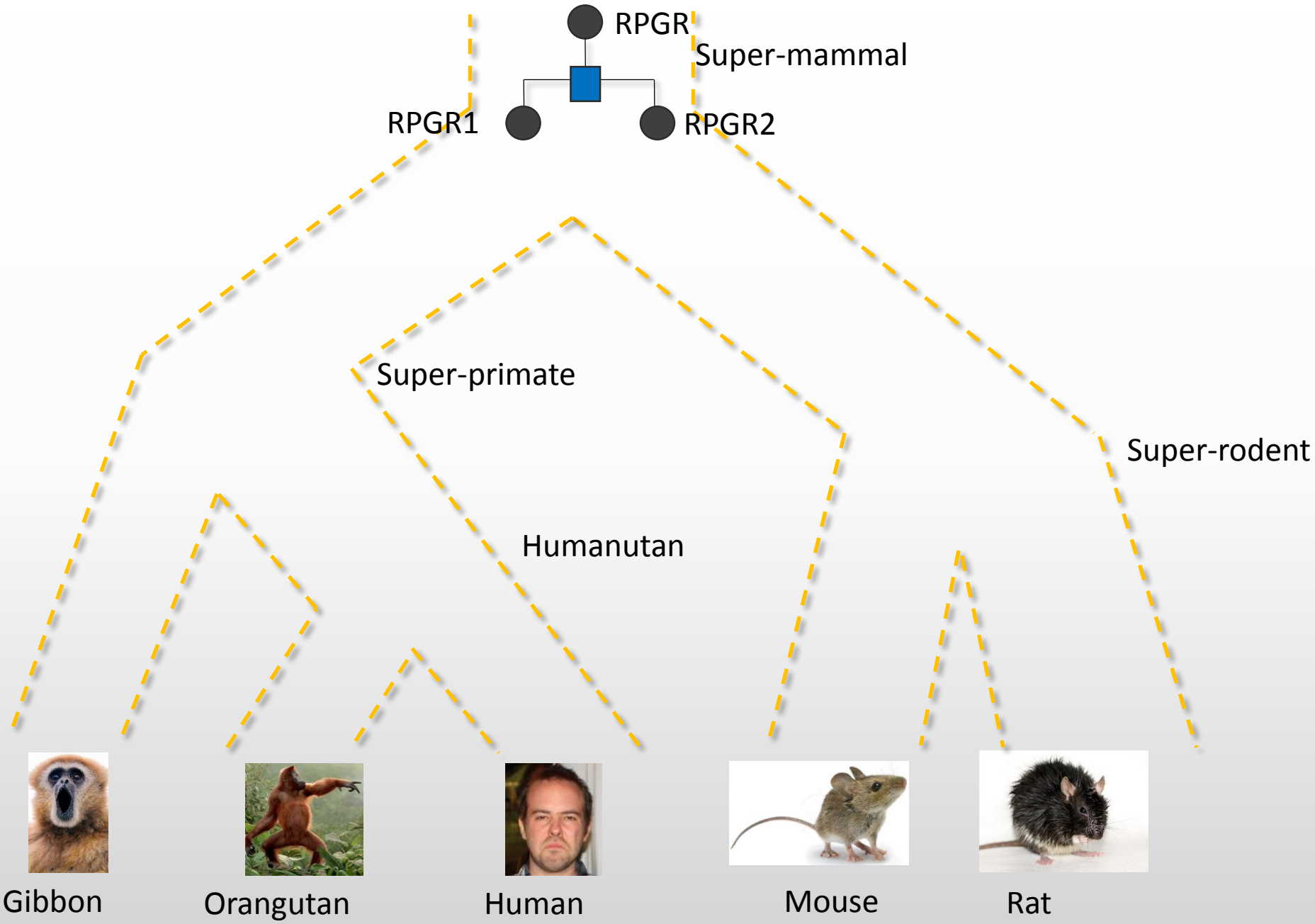
History = gene tree labeled with duplications and speciations

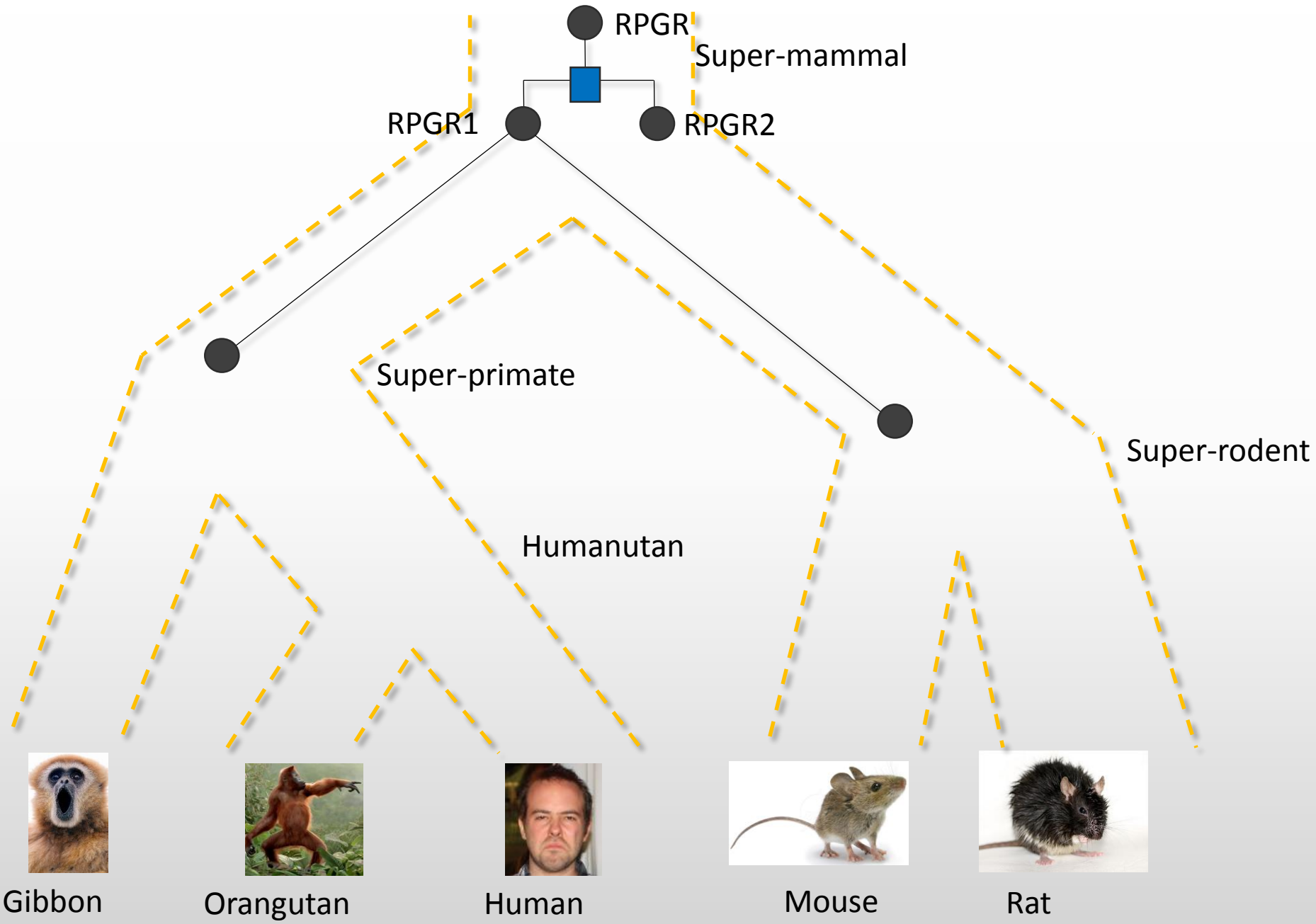


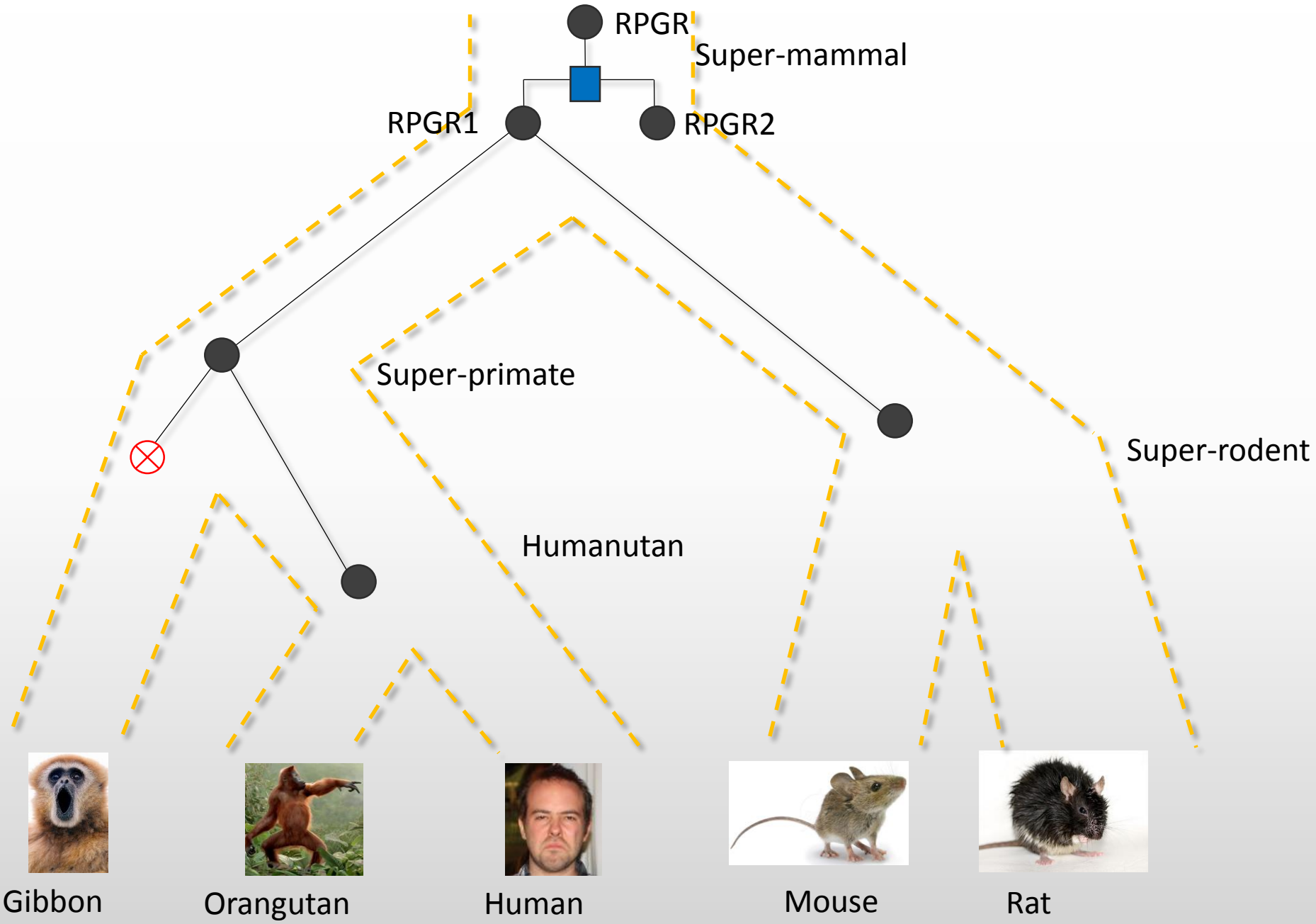

Super-mammal
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

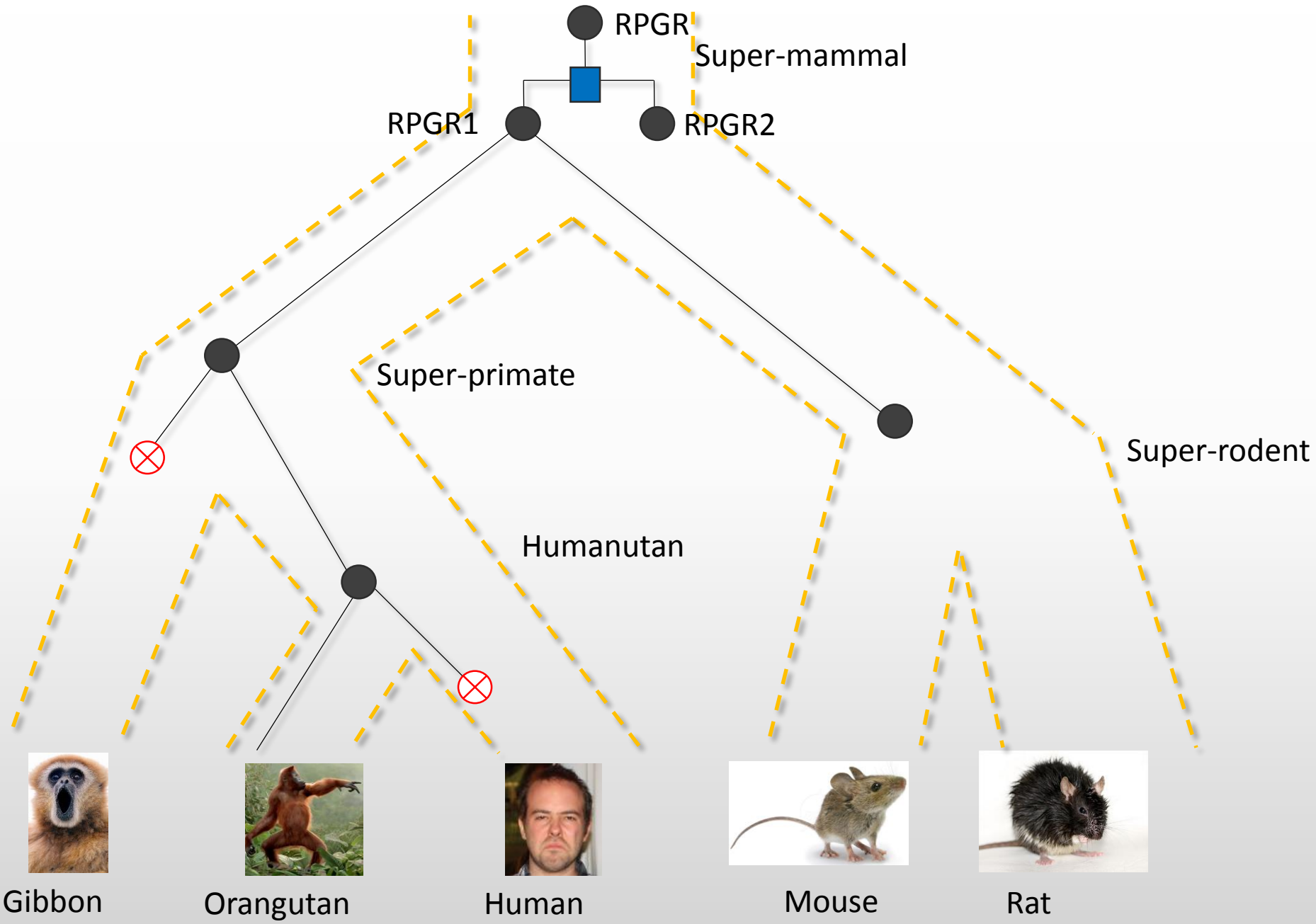

Super-mammal
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

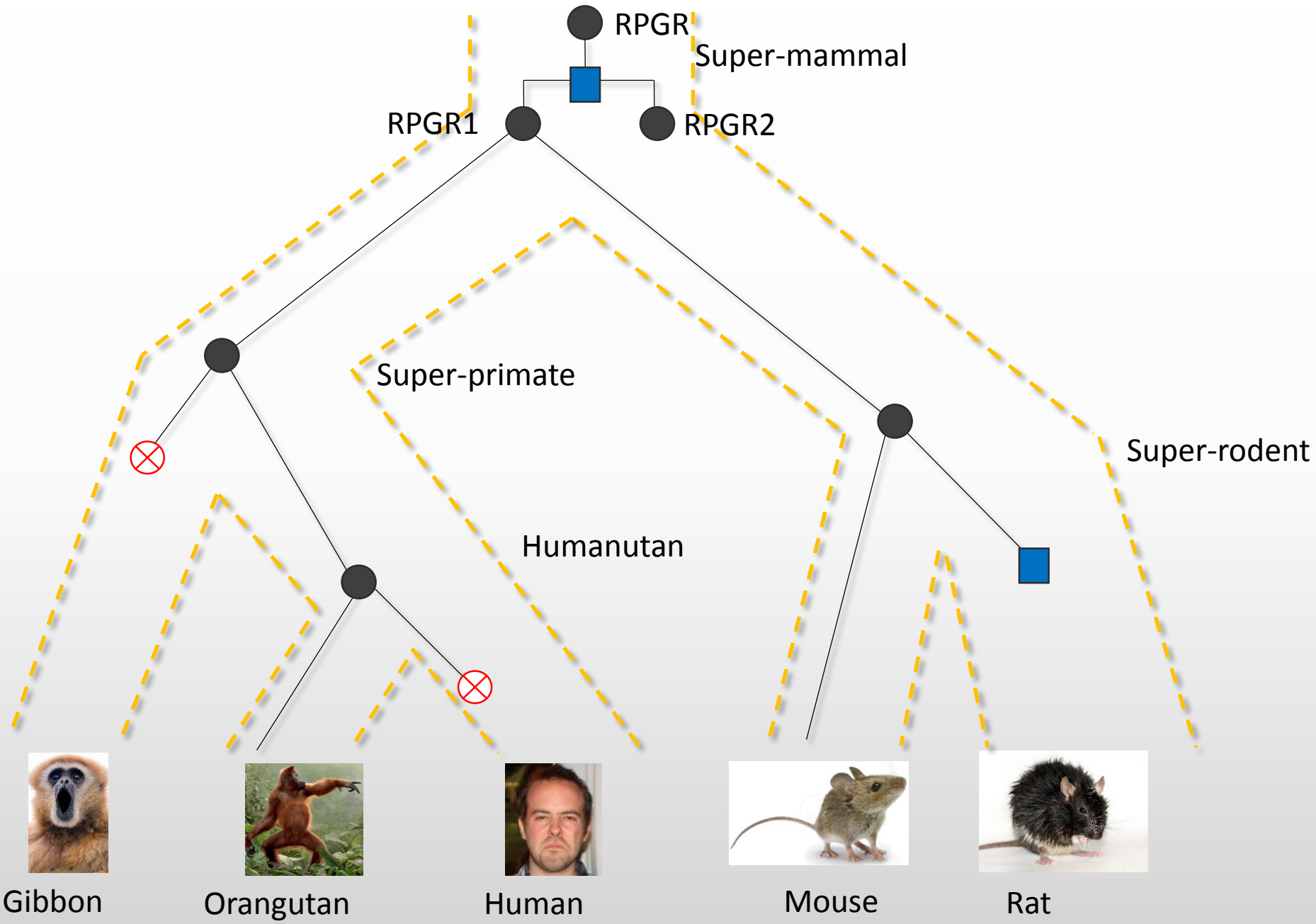

RPGR
Super-mammal
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

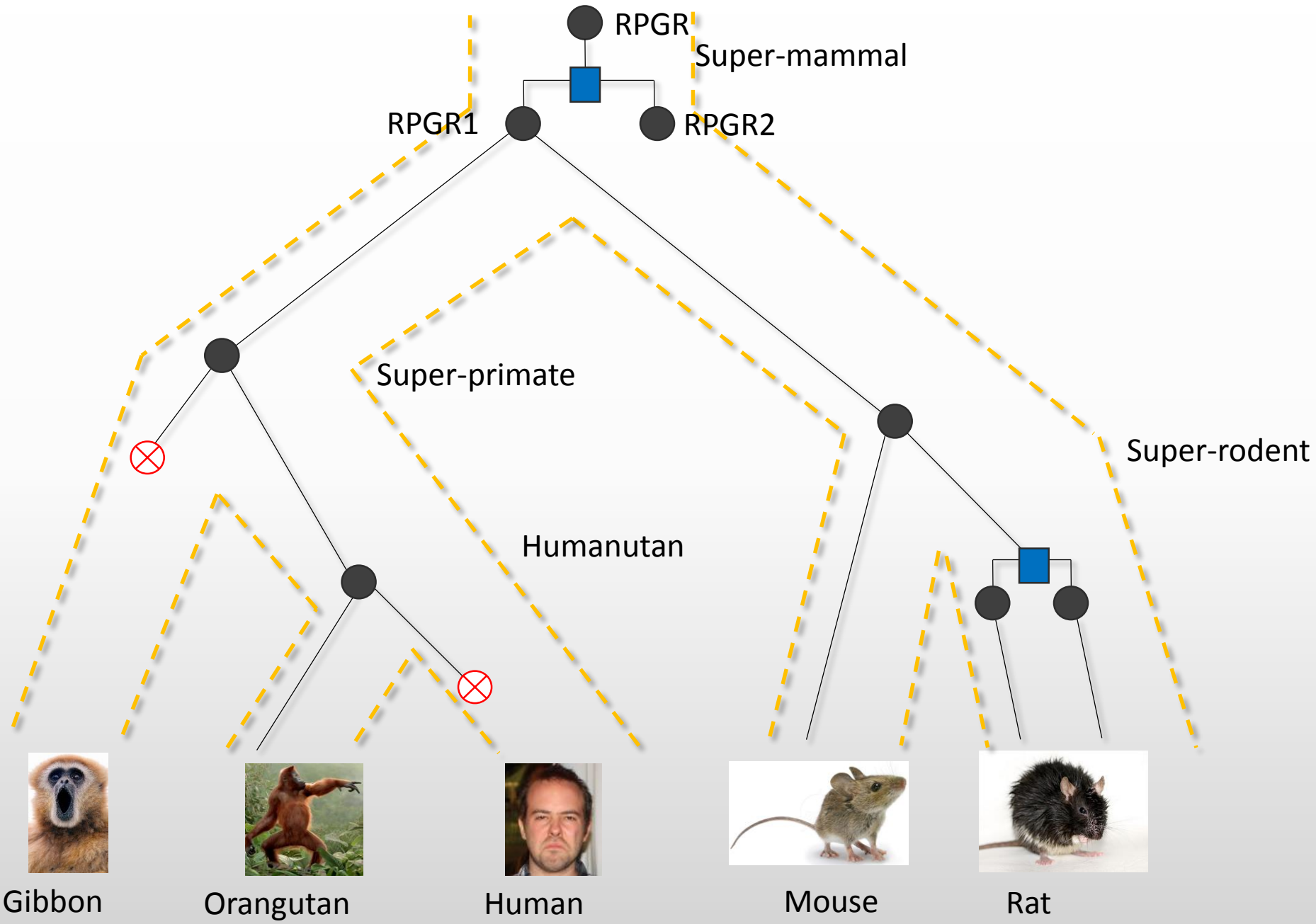

RPGR
Super-mammal
RPGR1
RPGR2
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

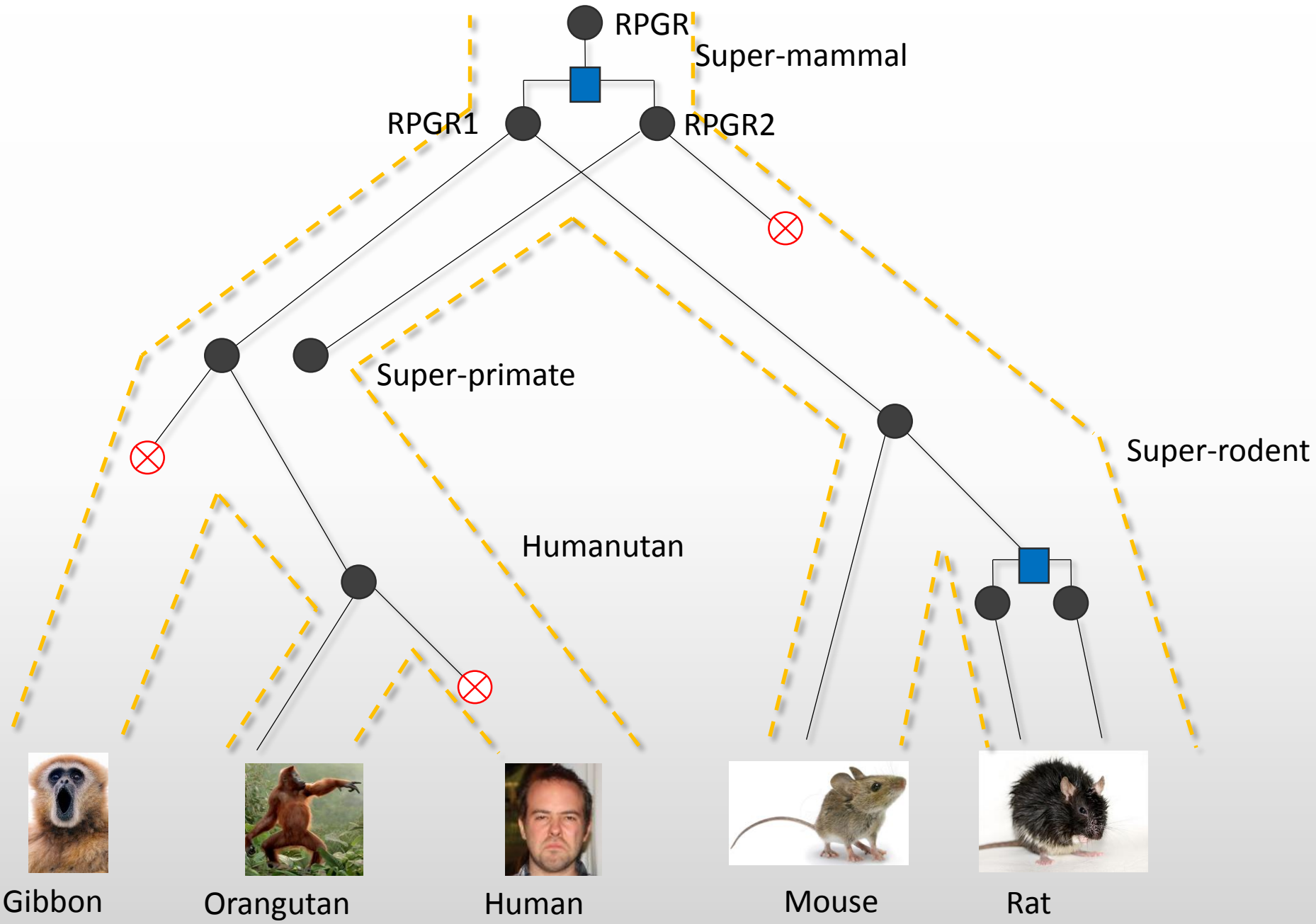

RPGR
Super-mammal
RPGR1
RPGR2
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

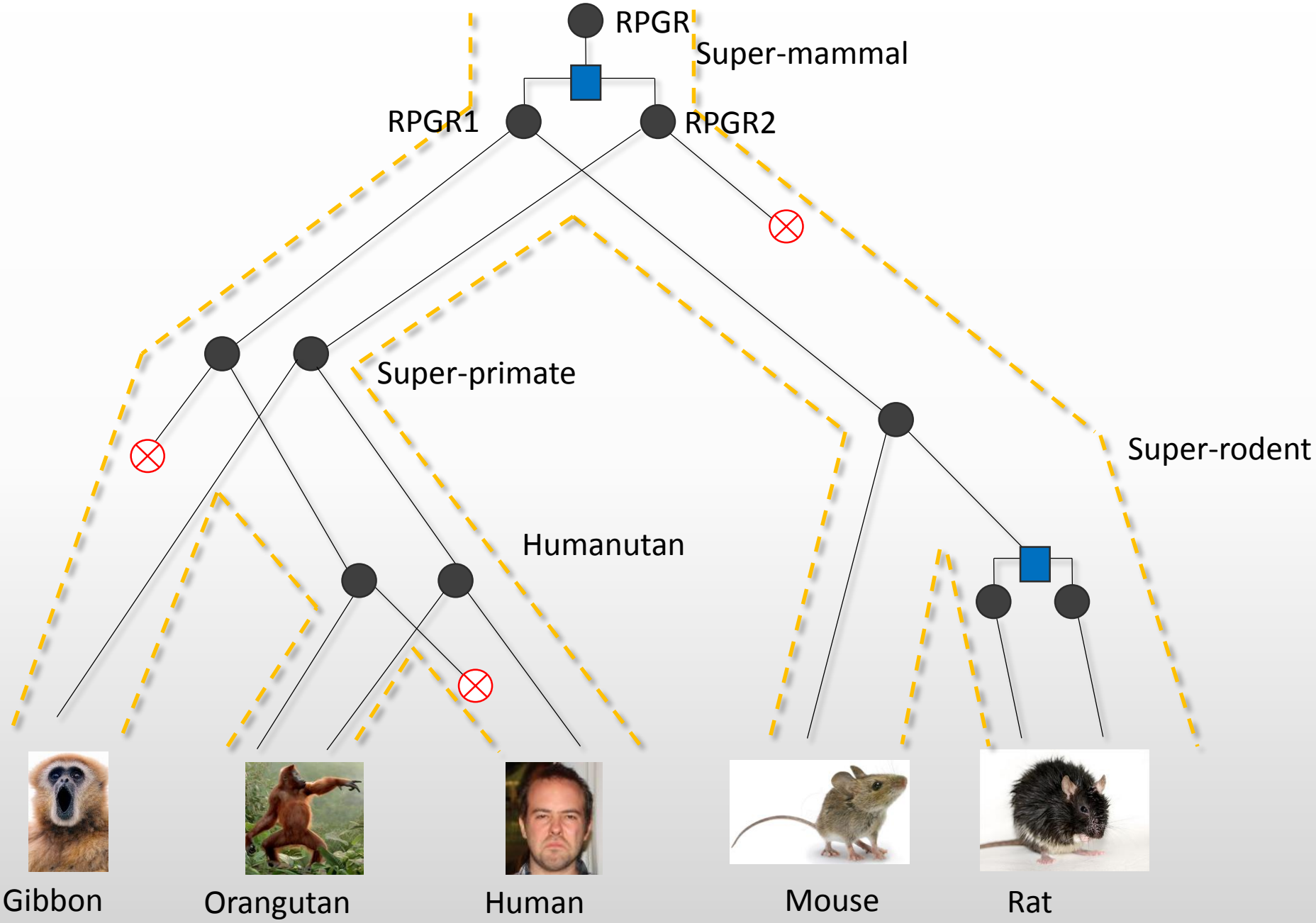

RPGR
Super-mammal
RPGR1
RPGR2
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

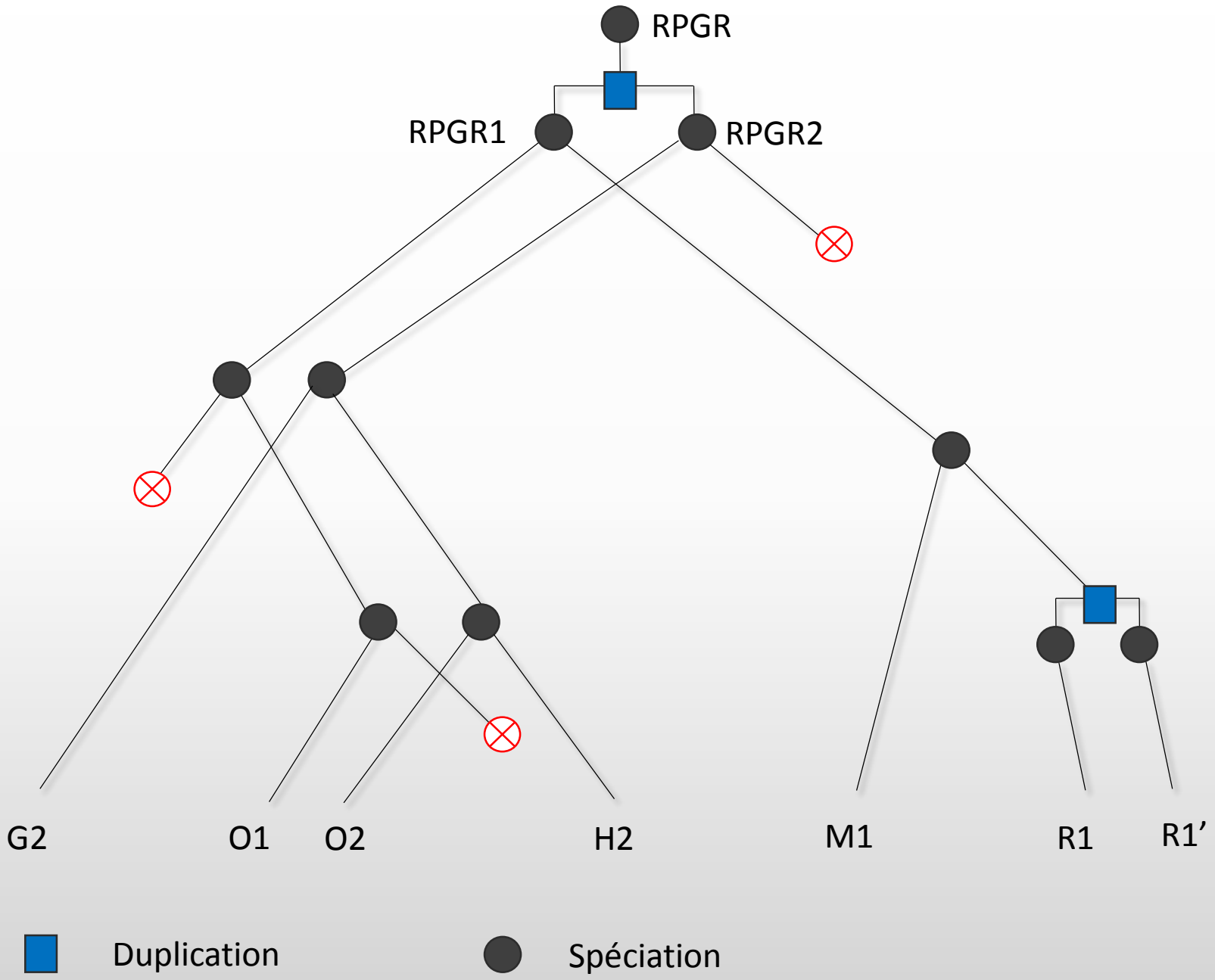

RPGR
Super-mammal
RPGR1
RPGR2
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

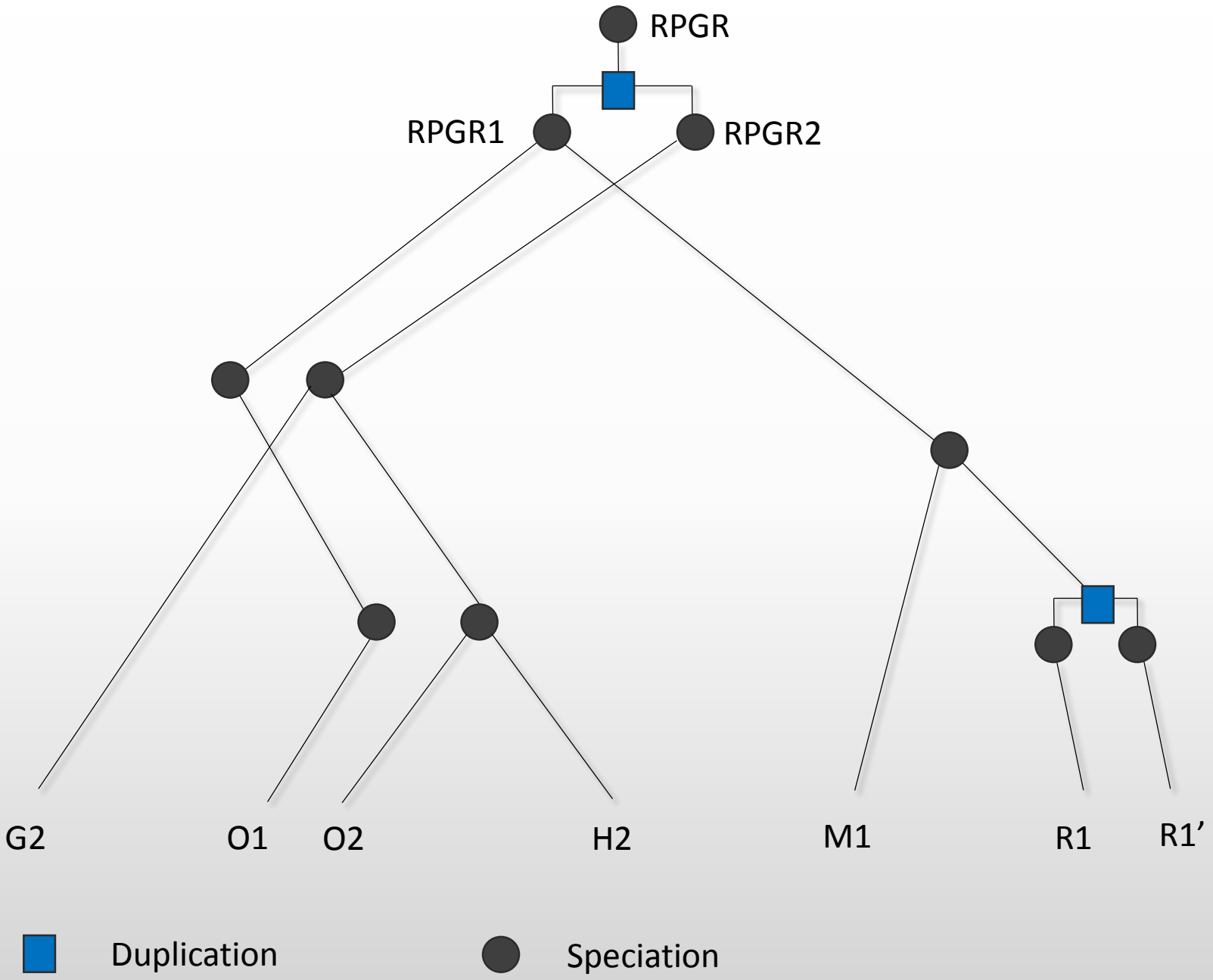

RPGR
Super-mammal
RPGR1
RPGR2
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

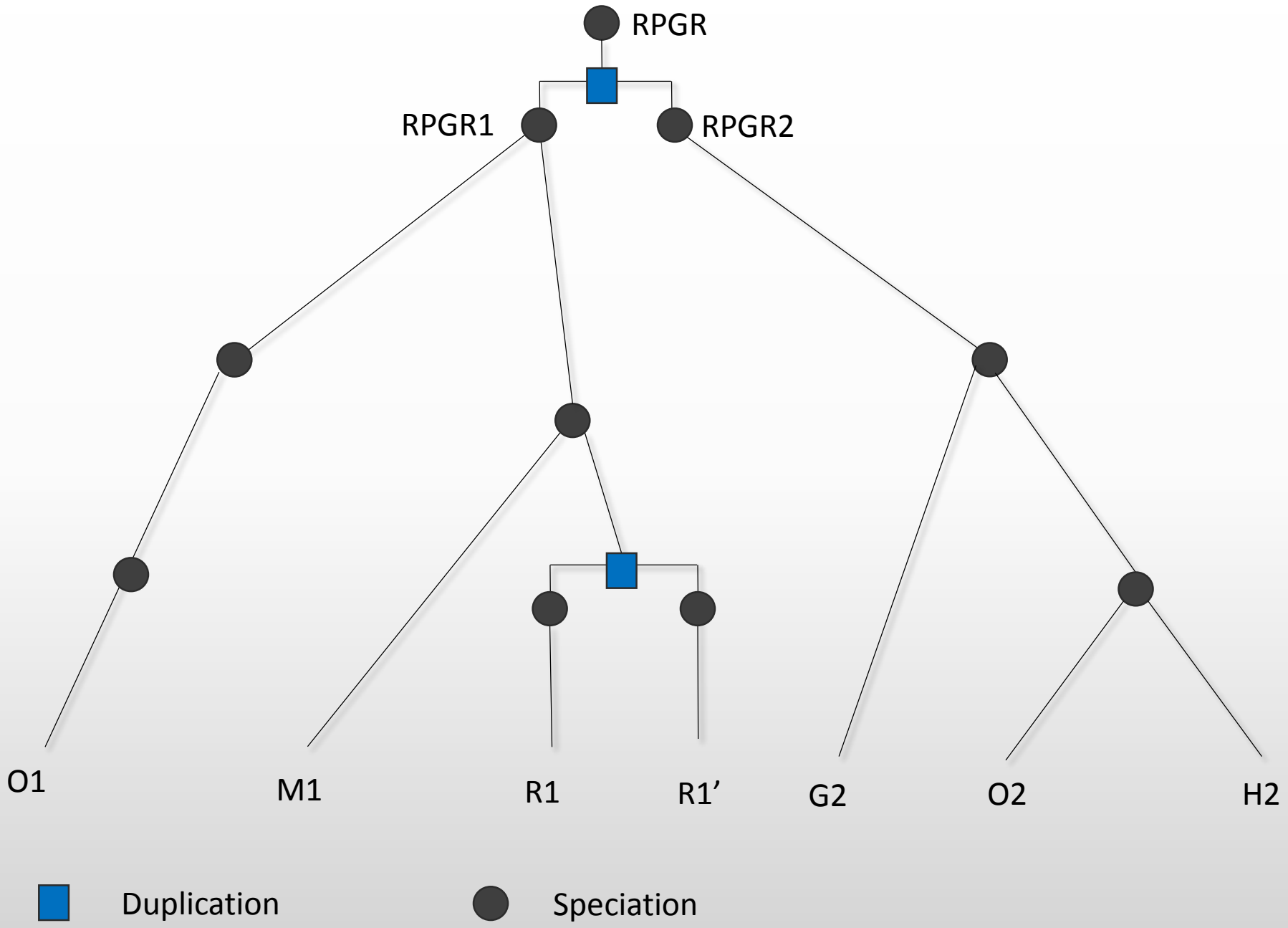

RPGR
Super-mammal
RPGR1
RPGR2
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

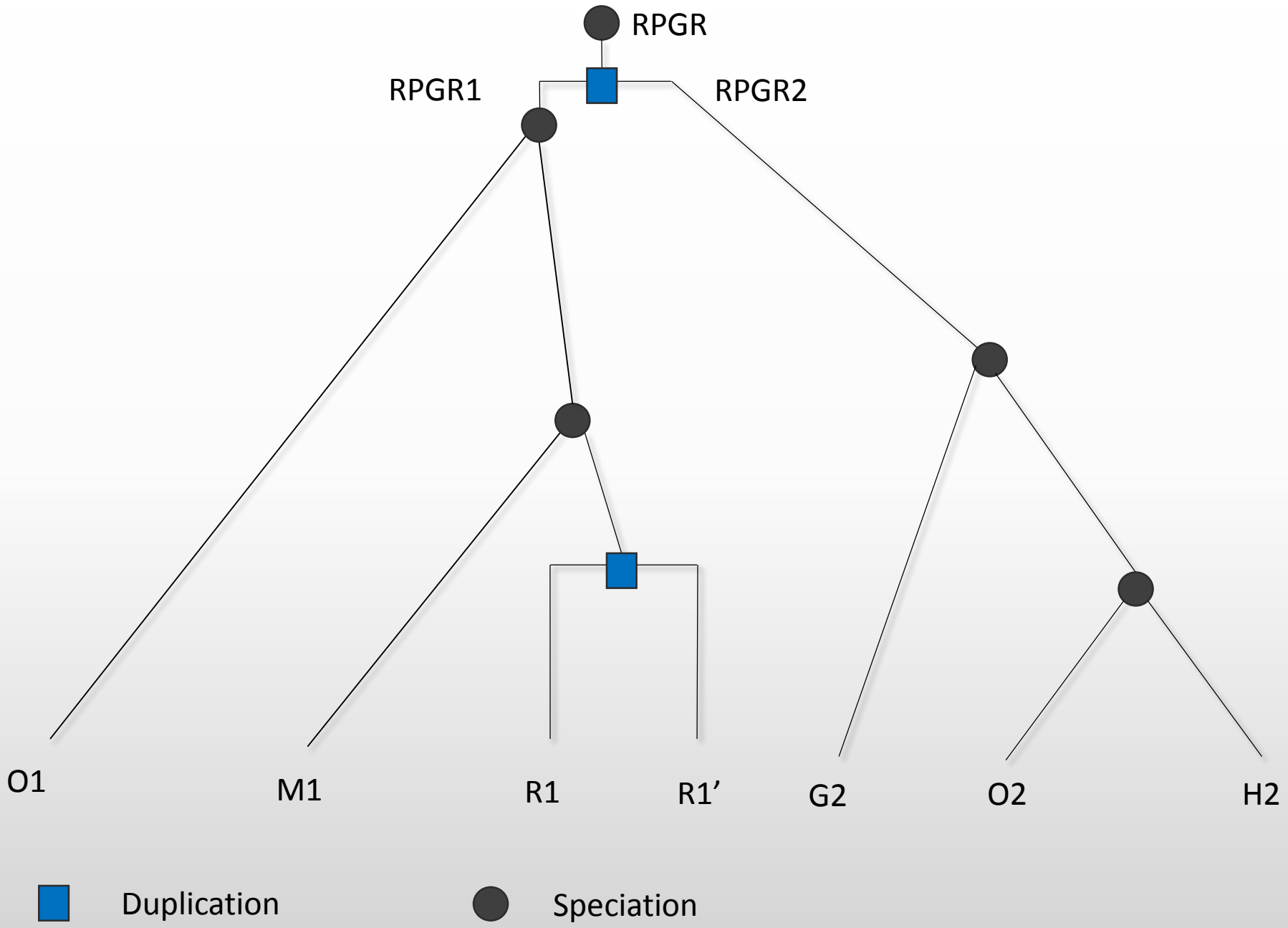

RPGR
Super-mammal
RPGR1
RPGR2
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat

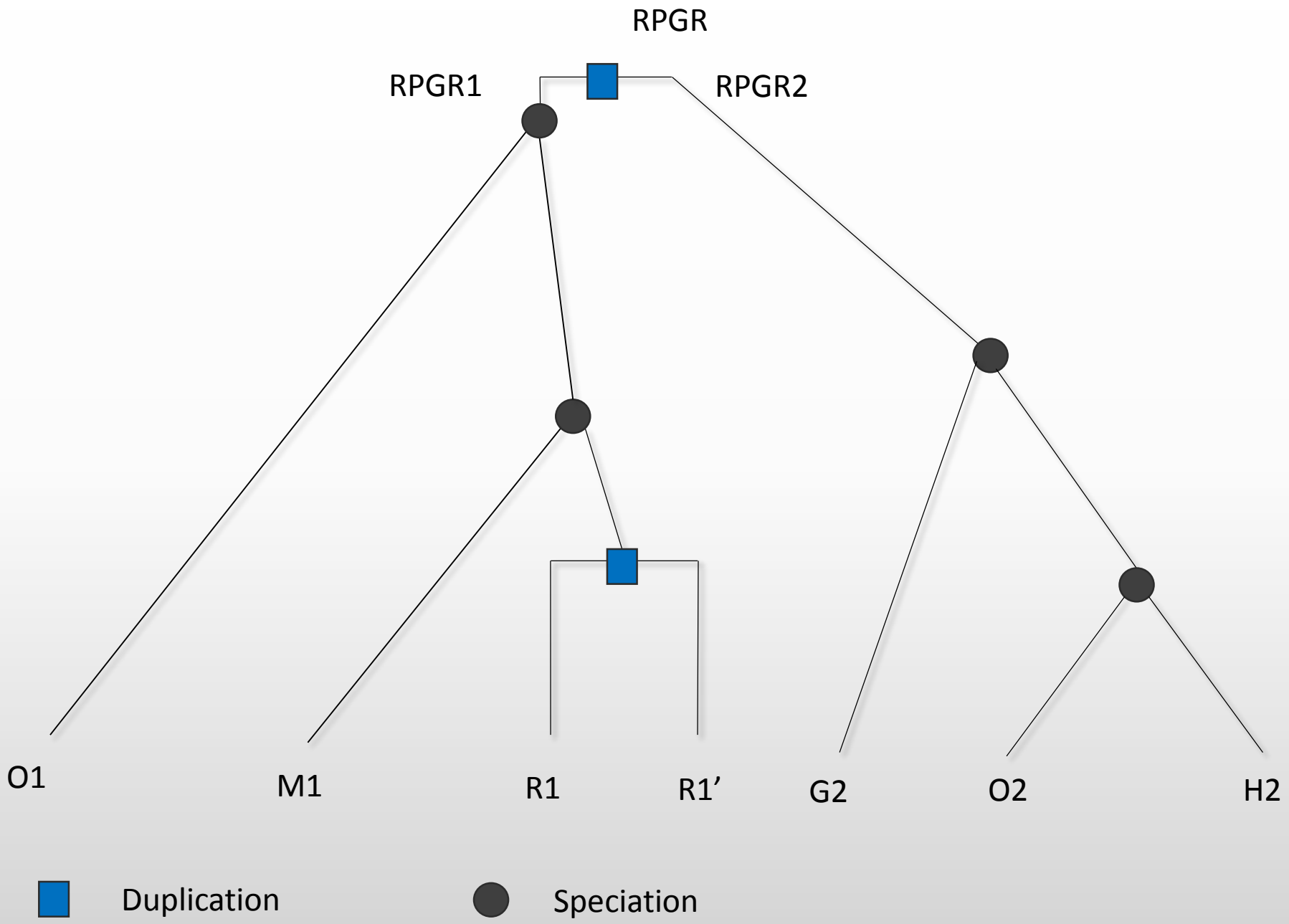

RPGR
Super-mammal
RPGR1
RPGR2
Super-primate
Super-rodent
Humanutan
Gibbon
Orangutan
Human
Mouse
Rat


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


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

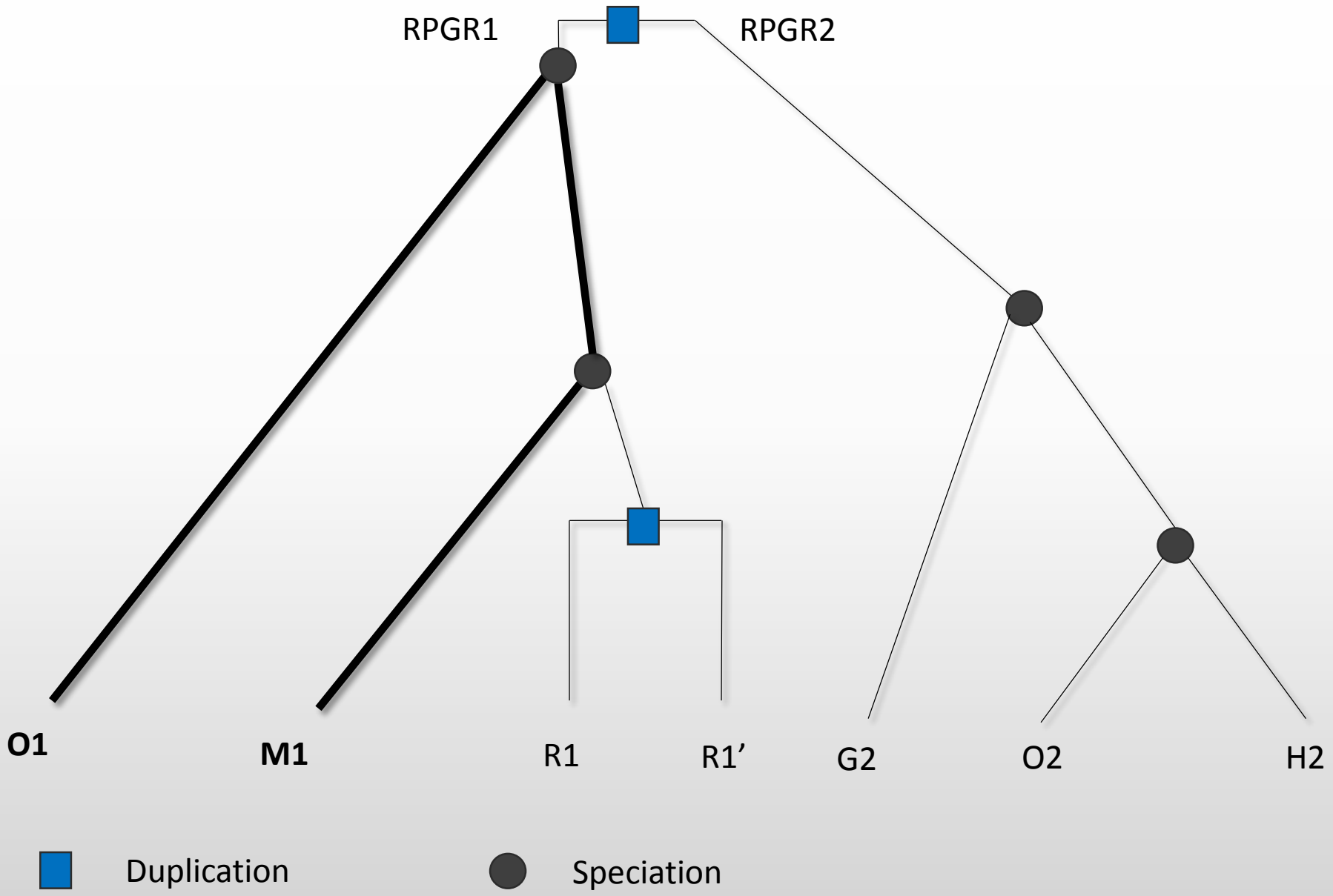

RPGR
RPGR1
RPGR2
O1
M1
R1
R1'
G2
O2
H2
Duplication
Speciation

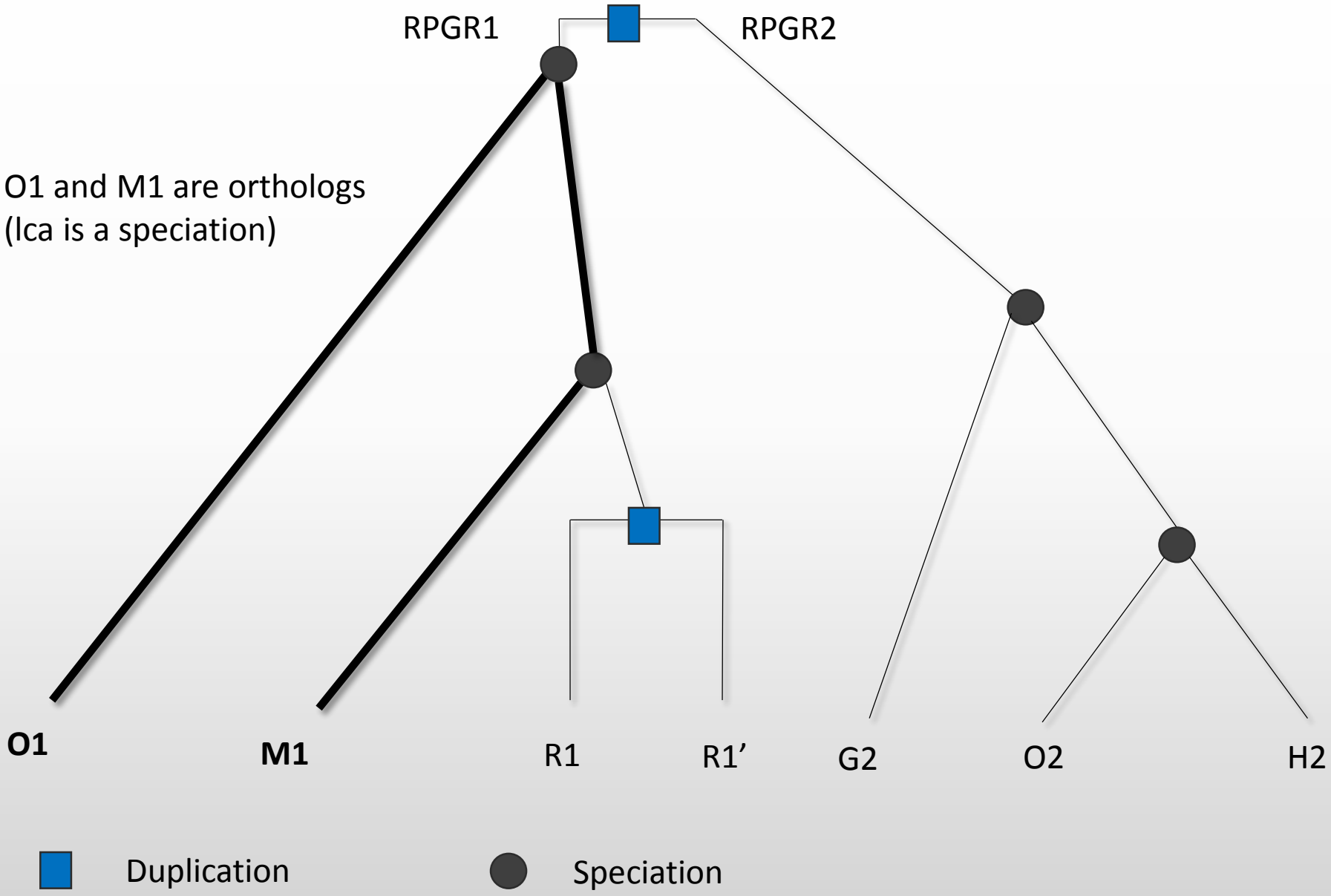
Orthologs et paralogs

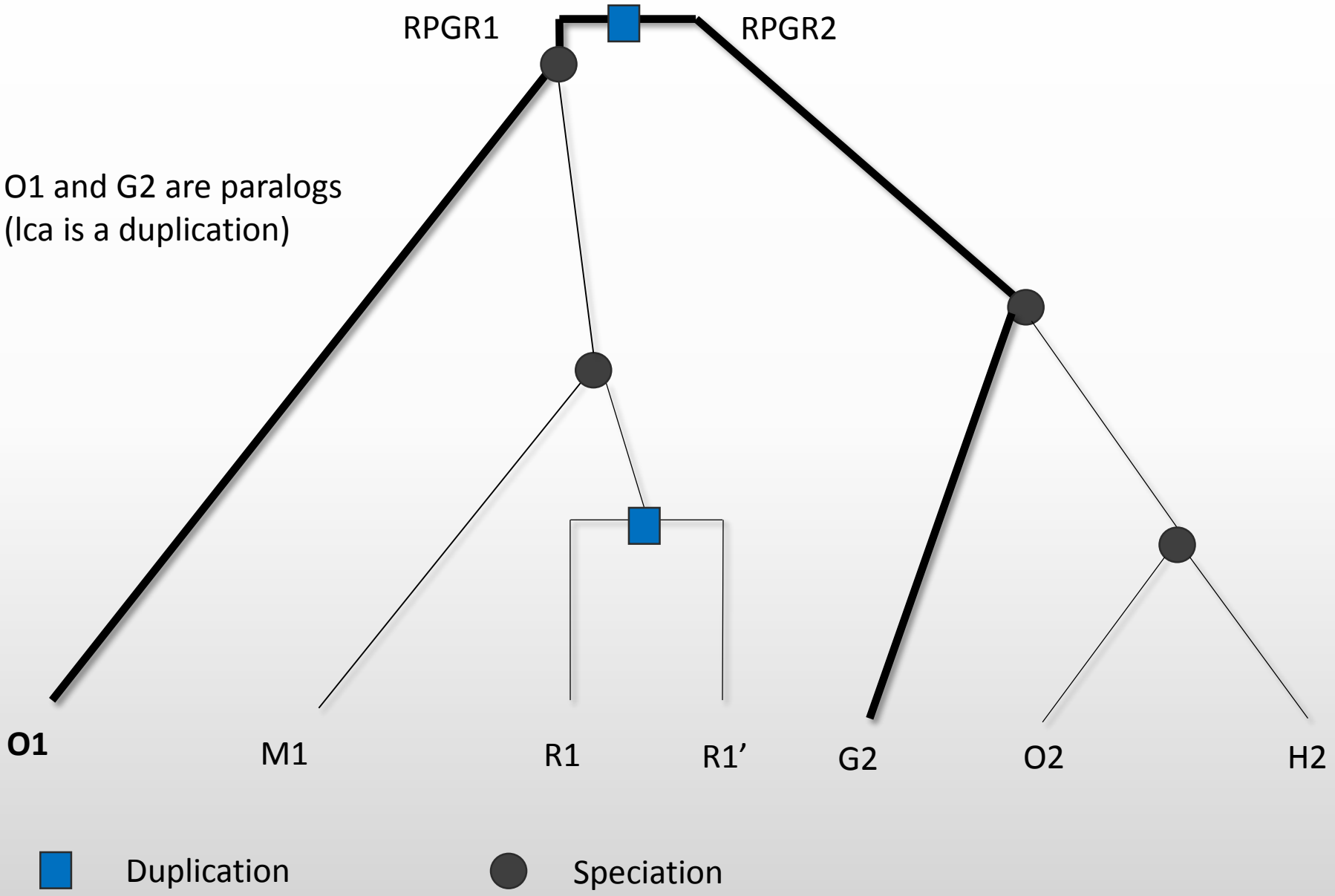
Two genes are:

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

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


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


RPGR1
RPGR2
O1 and M1 are orthologs
(lca is a speciation)
O1
M1
R1
R1'
G2
O2
H2
Duplication
Speciation


RPGR1
RPGR2
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 to share the same functionality

(e.g. COG, eggNOG databases)

Why bother?

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

- Any hope of proving or disproving this conjecture first requires computational tools that can accurately infer gene relations.

Why bother?

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

- Any hope of proving or disproving this conjecture first requires computational tools that can accurately infer gene relations.

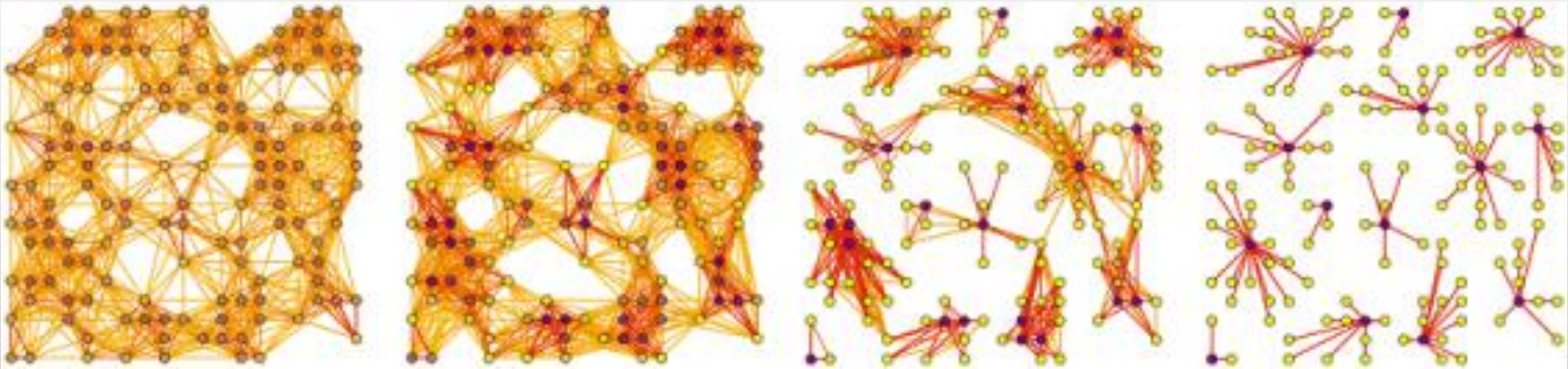
Quest For Orthologs consortium: "a joint effort to benchmark, improve and standardize orthology predictions through collaboration, the use of shared reference datasets, and evaluation of emerging new methods".

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

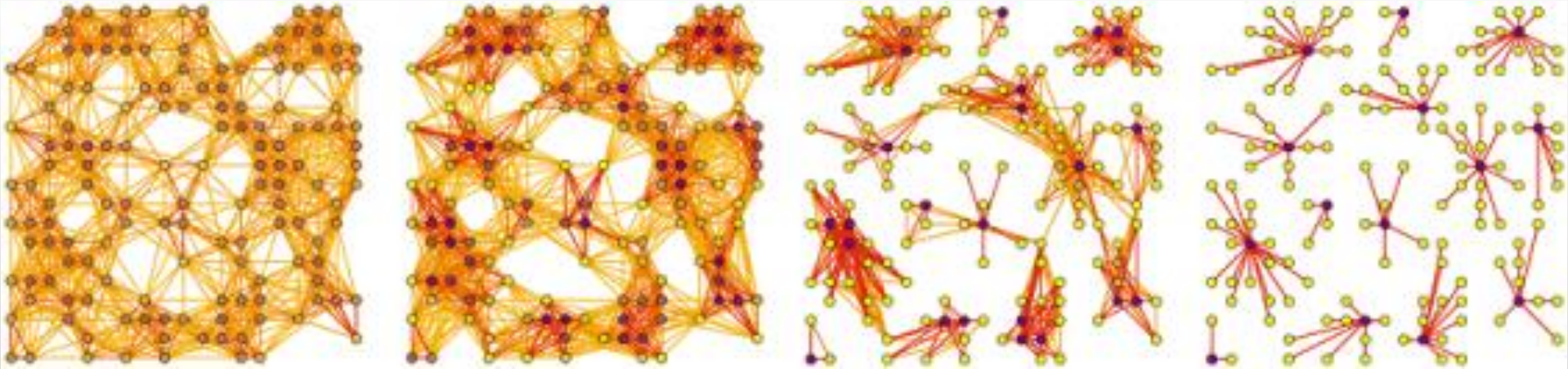
  Otherwise, too many false positives occur
- OrthoMCL, InParanoid, proteinortho, …



Traditional inference method

These methods are very often **incomplete** - have false positives or false negatives.

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



What we want to do

Given a set of orthologs / paralogs:

- 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 a minimal way

  Everything is NP-Complete

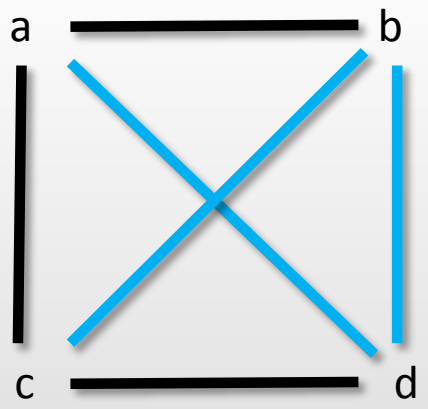
  Approximation algorithms

Validation of gene relations

Orthology/paralogy graph

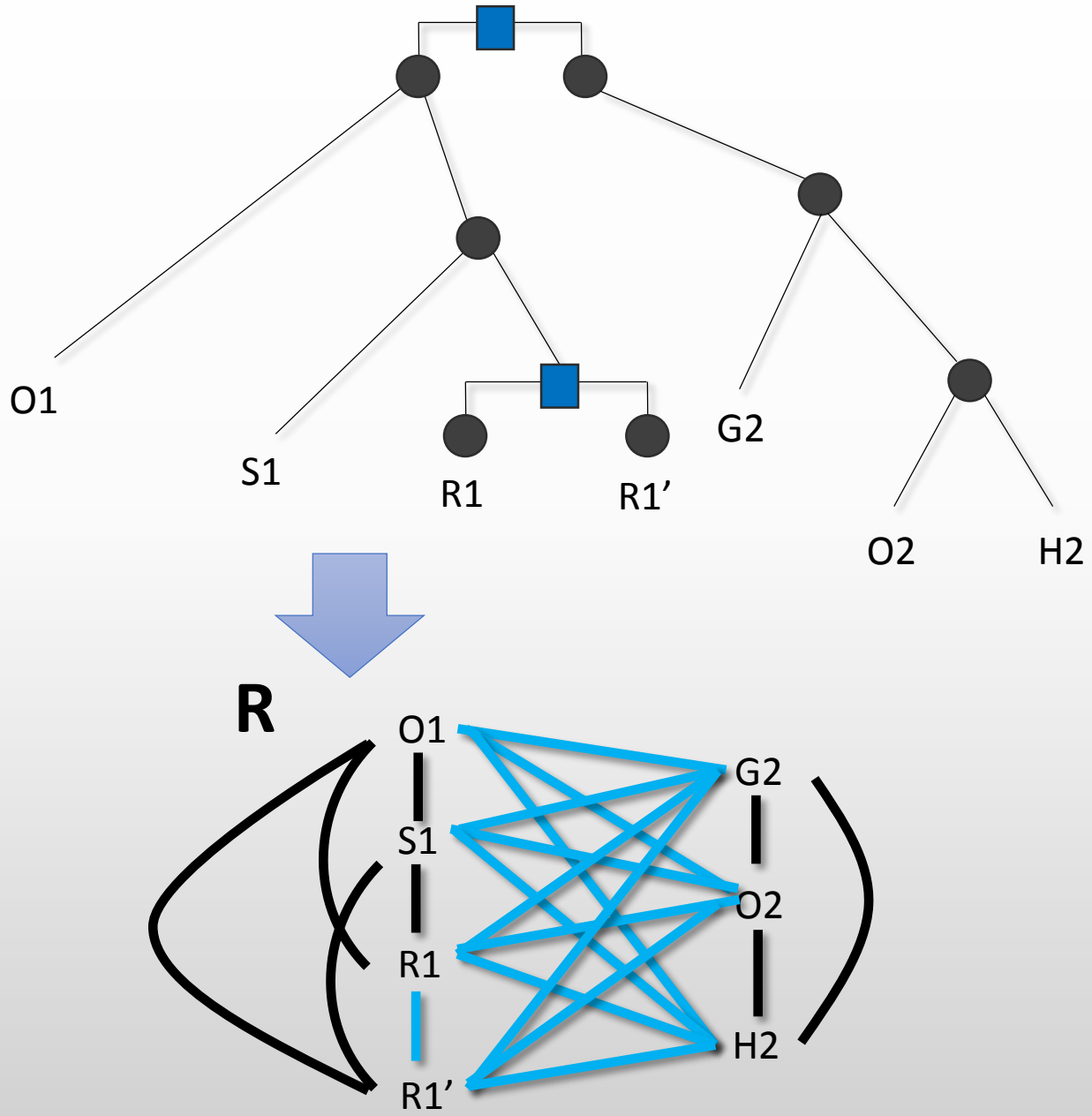
Orthologs = (a,b) (a, c) (c, d)

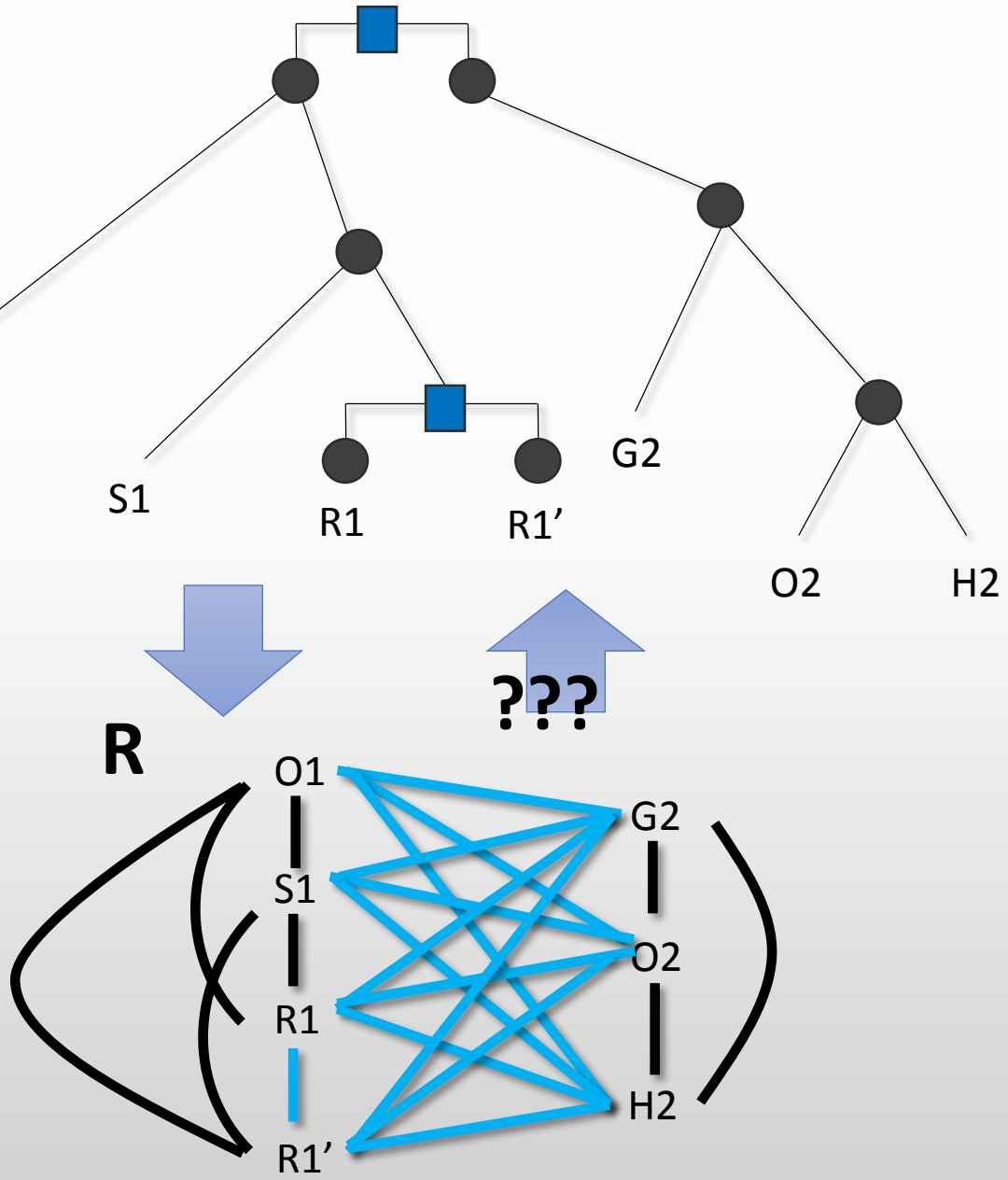
Paralogs = (a, d) (b, c) (b, d)

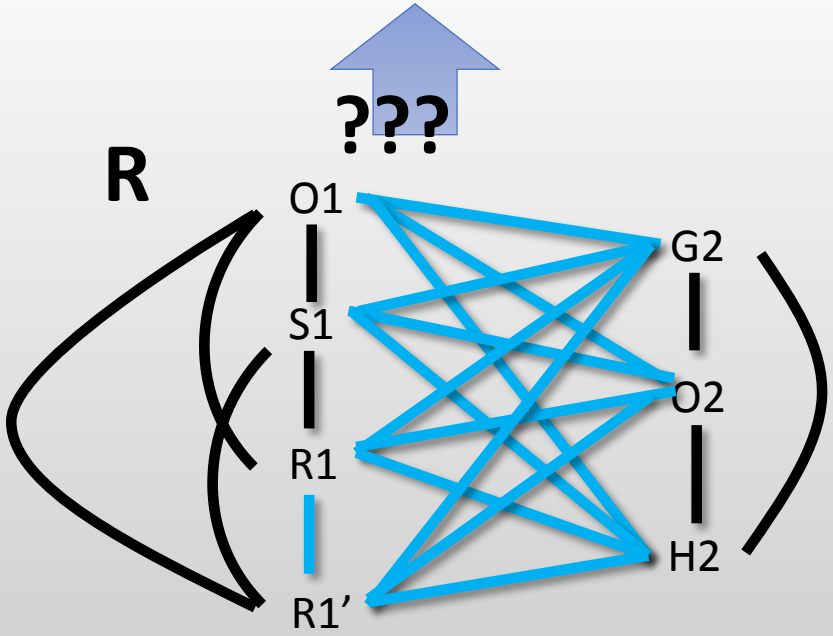


Orthologs

Paralogs


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

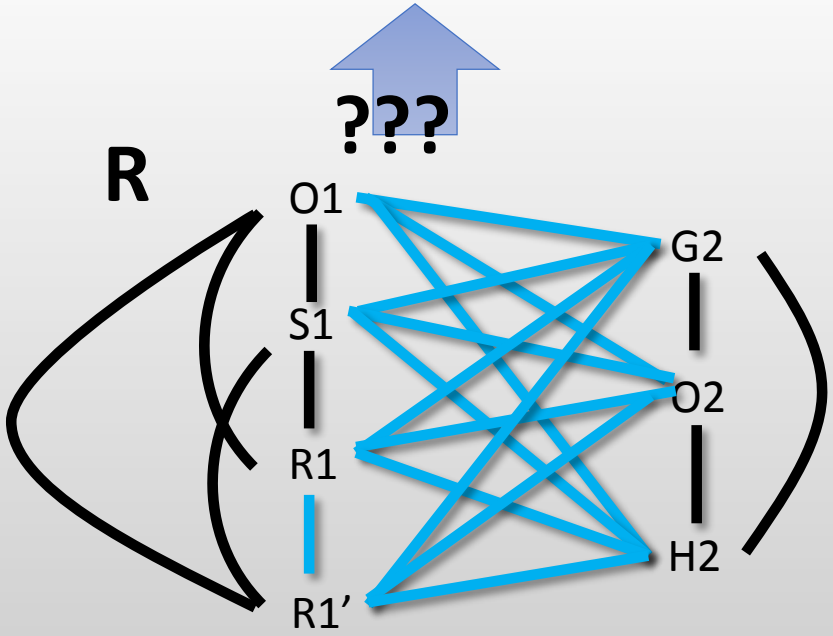

O1
S1
R1
R1'
G2
O2
H2
???
R
O1
S1
R1
R1'
G2
O2
H2


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

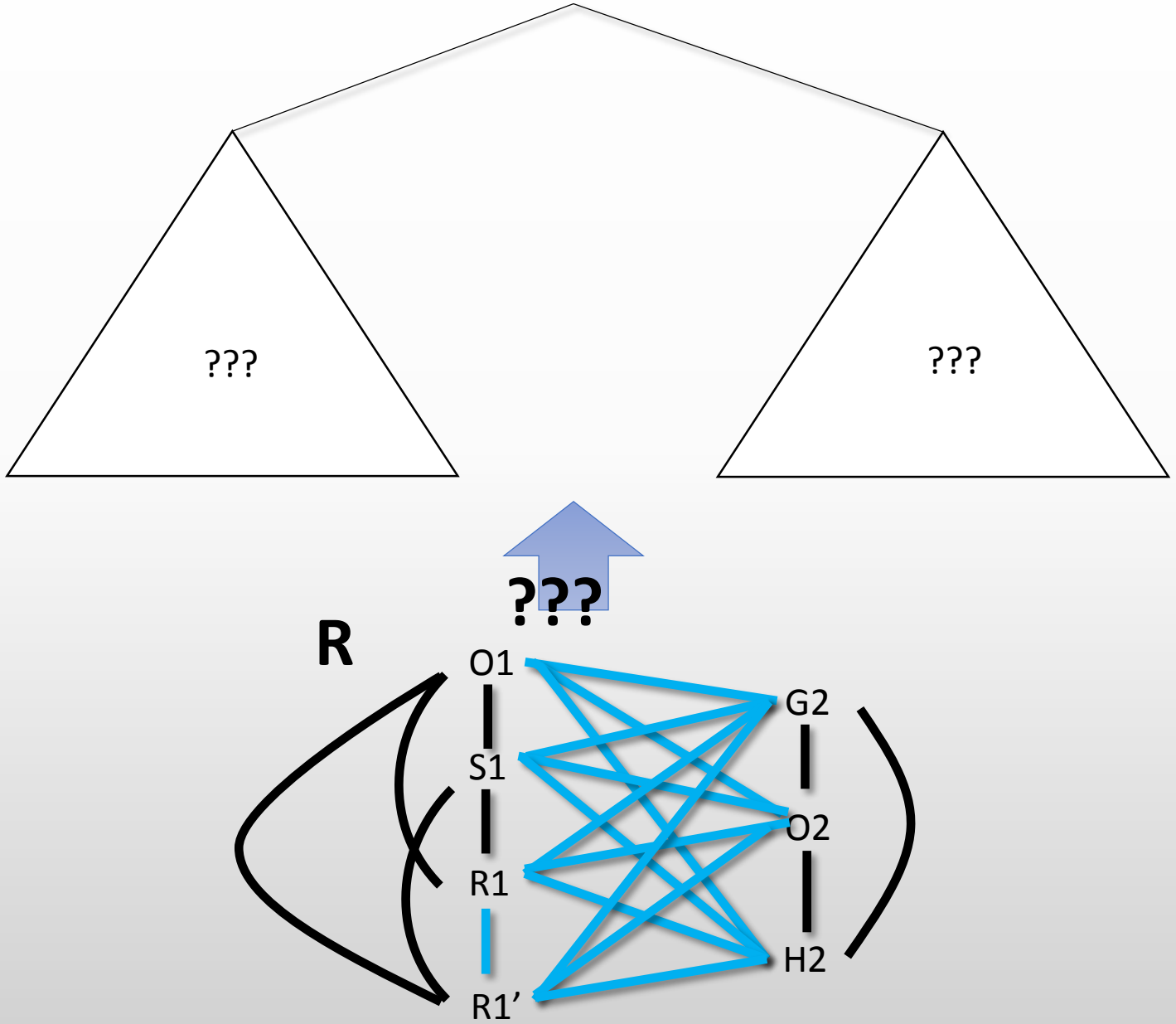
Problem :

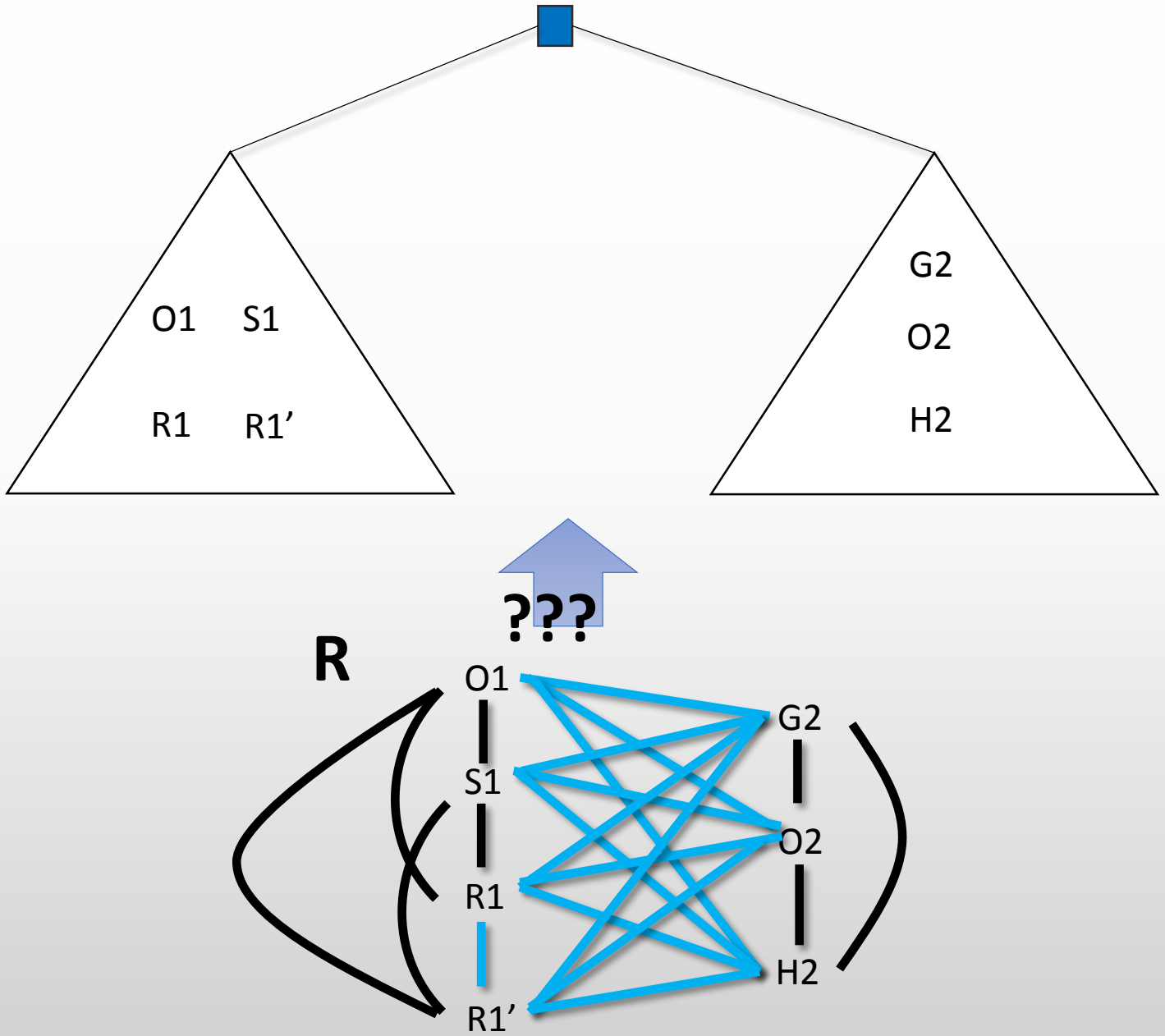
Given a relation graph R, is R **satisfiable**?

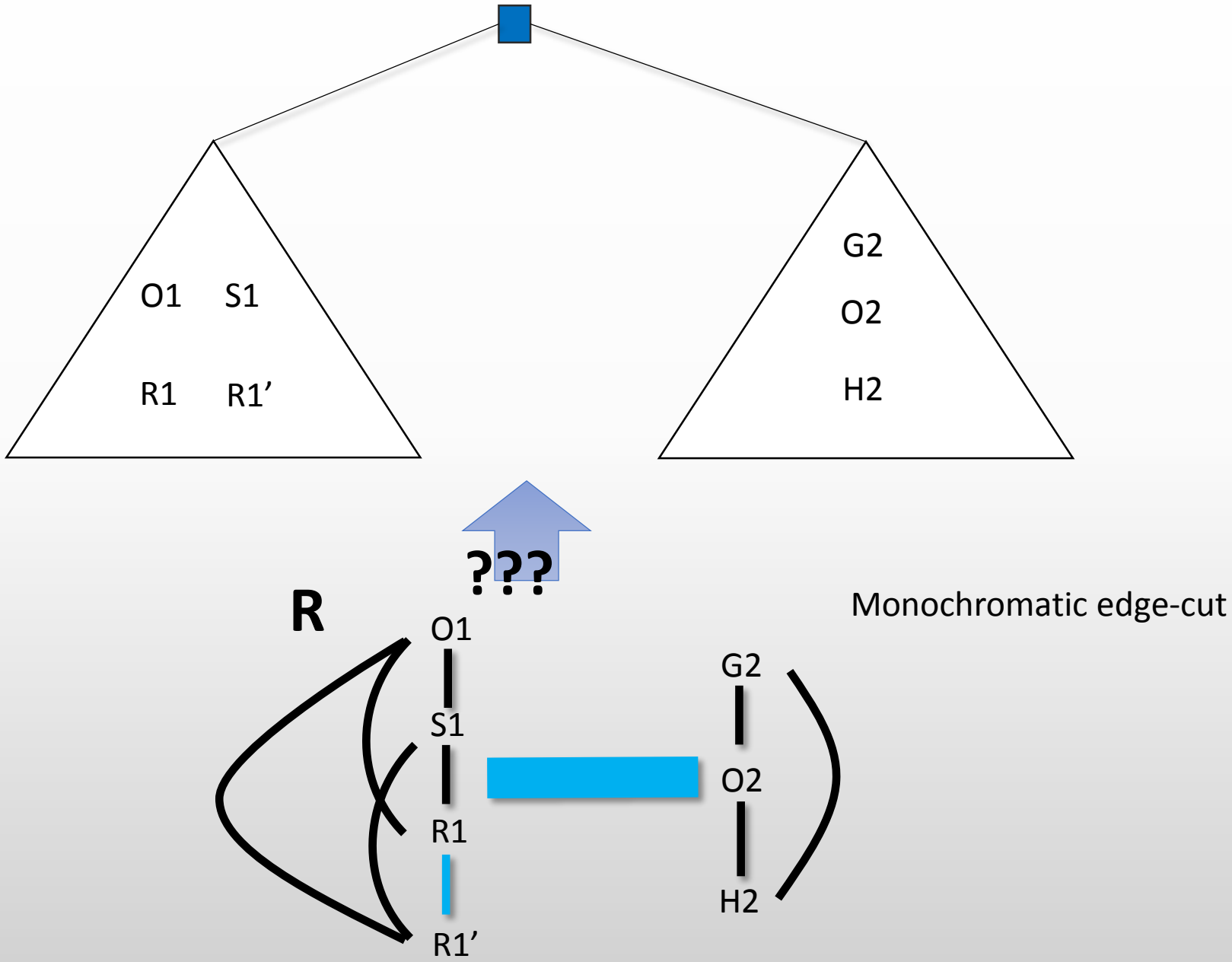
Does there exist a gene tree G that display the relations of R ?

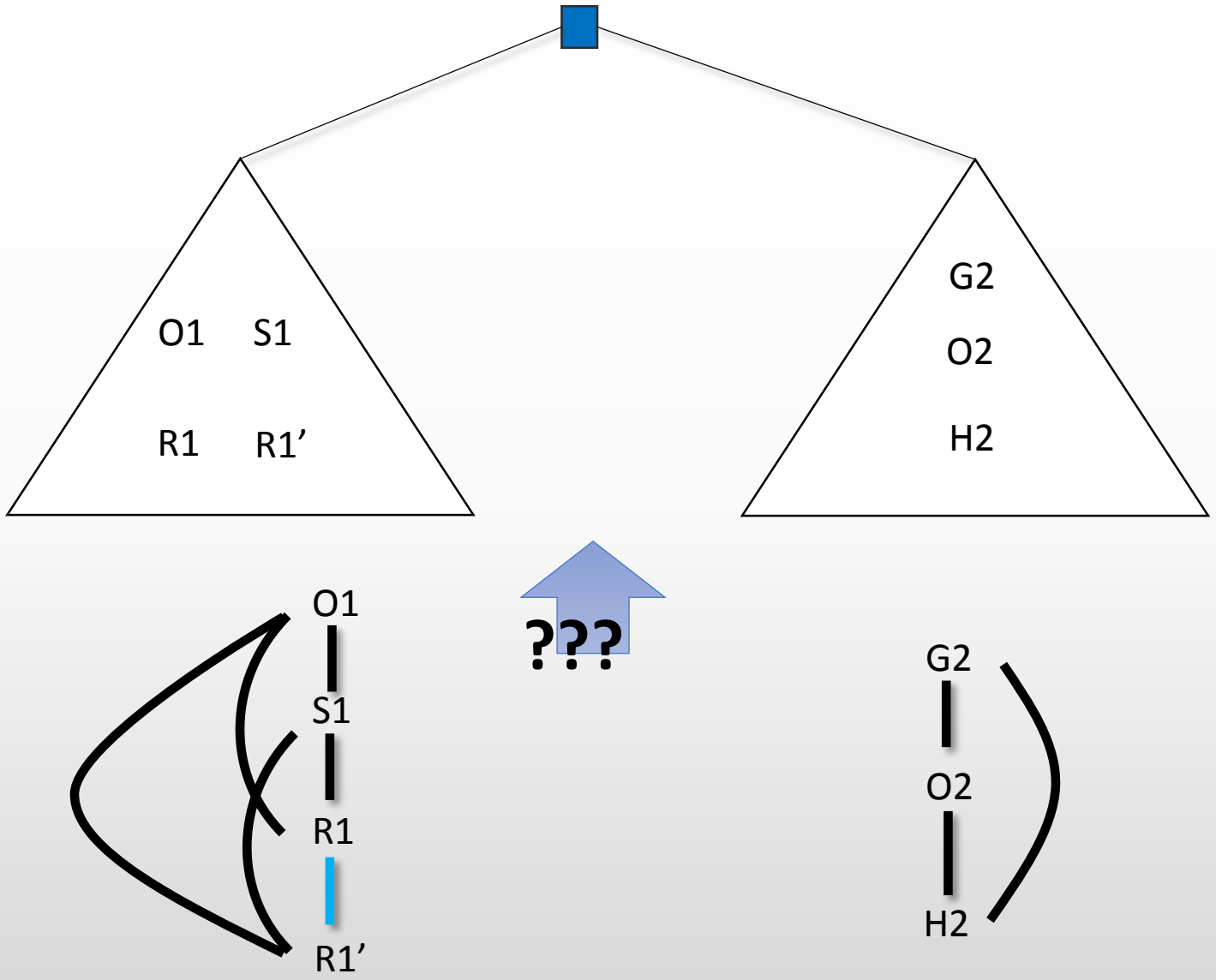


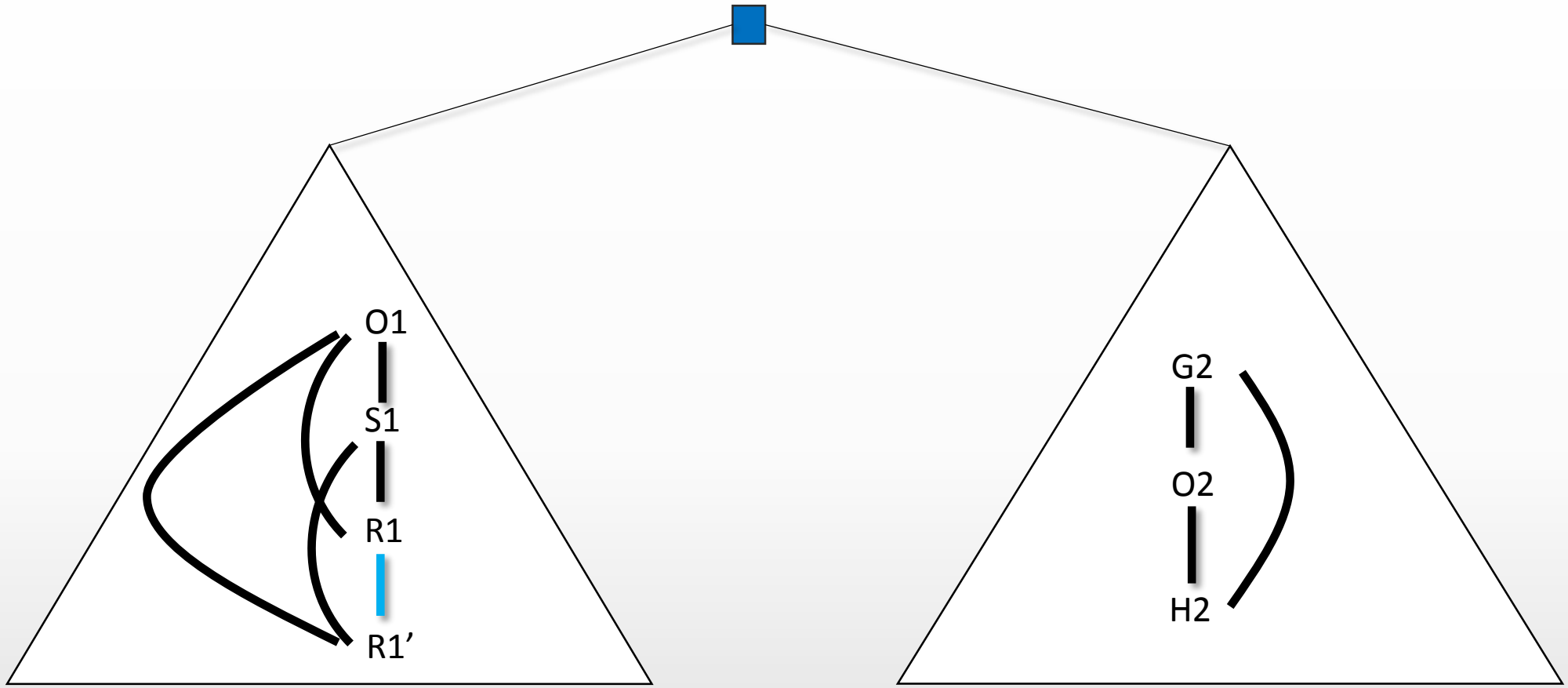
Let's say it exists…what is the first split then ?

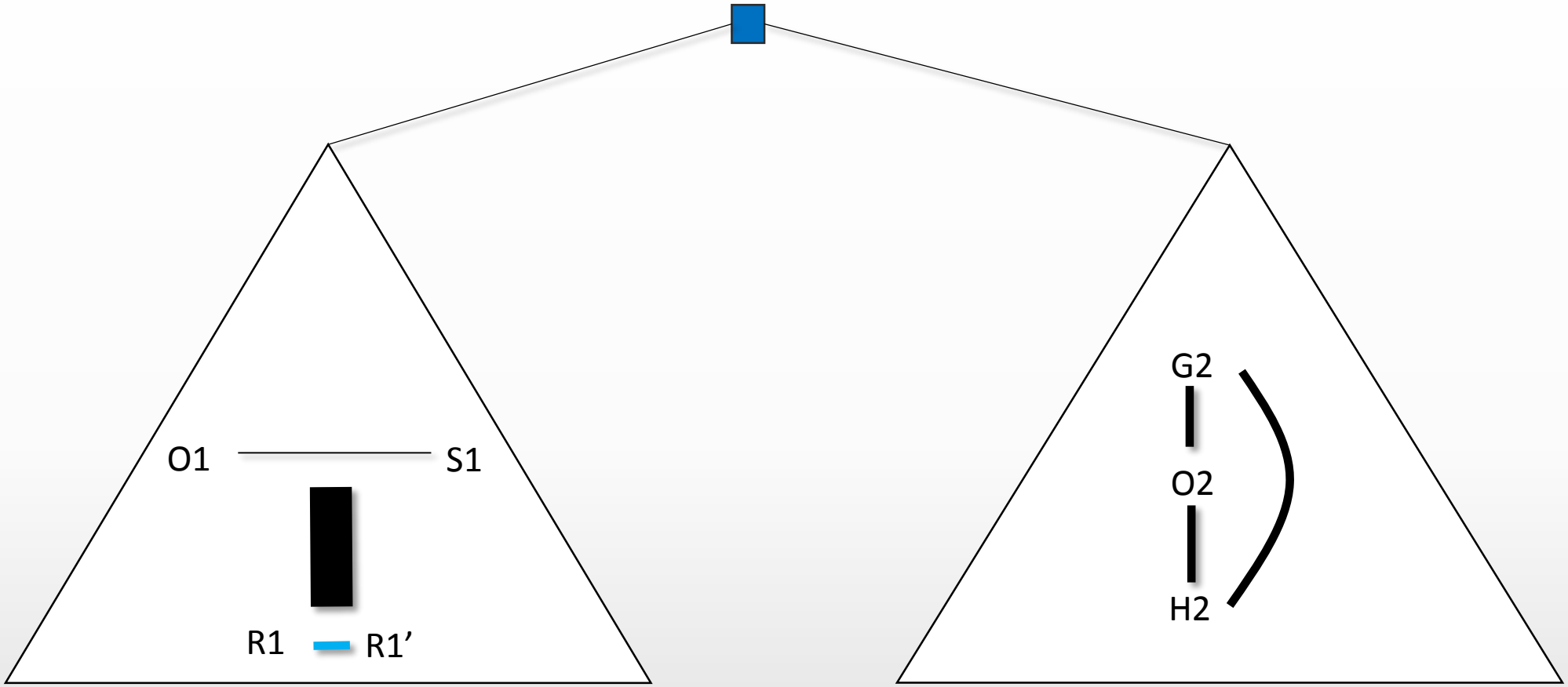


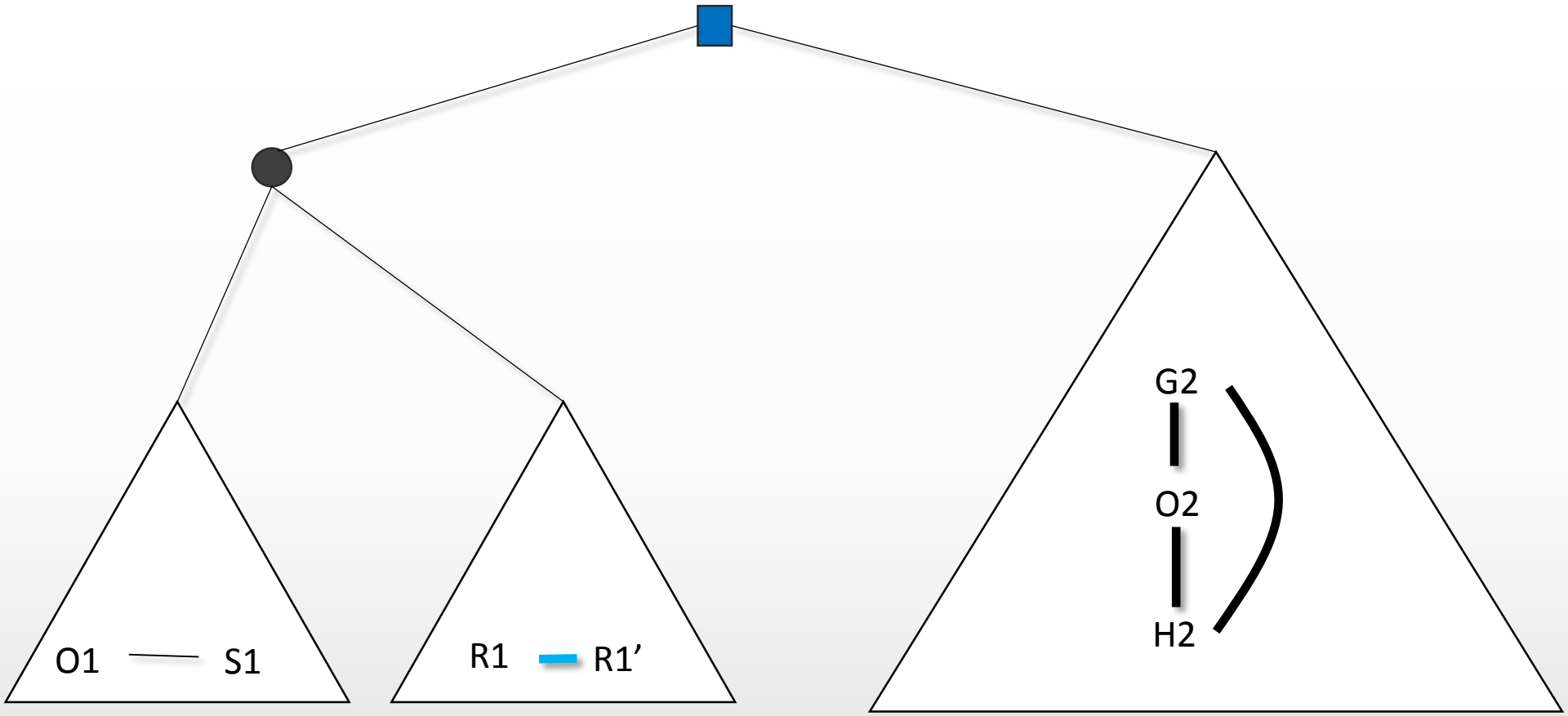

O1
S1
R1
R1'
G2
O2
H2
???
R
O1
S1
R1
R1'
G2
O2
H2


O1
S1
R1
R1'
G2
O2
H2
???
R
O1
S1
R1
R1'
G2
O2
H2
Monochromatic edge-cut


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


O1
S1
R1
R1’
G2
O2
H2

Lemma:

If each subgraph of the relation graph R has a **monochromatic edge-cut**, we can build a gene tree from R.

Conversely??

If R has a subgraph with no such cut, does it mean that we can't build a gene tree?

Lemma:

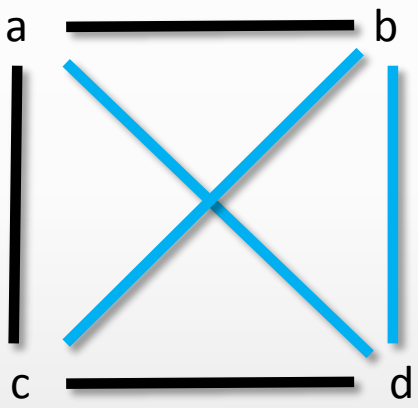
If each subgraph of the relation graph R has a **monochromatic edge-cut**, we can build a gene tree from R.

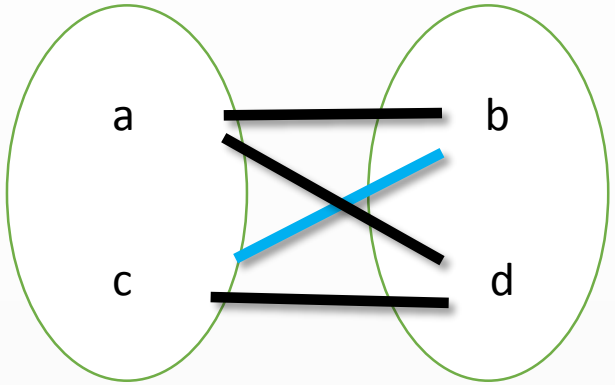
Conversely??

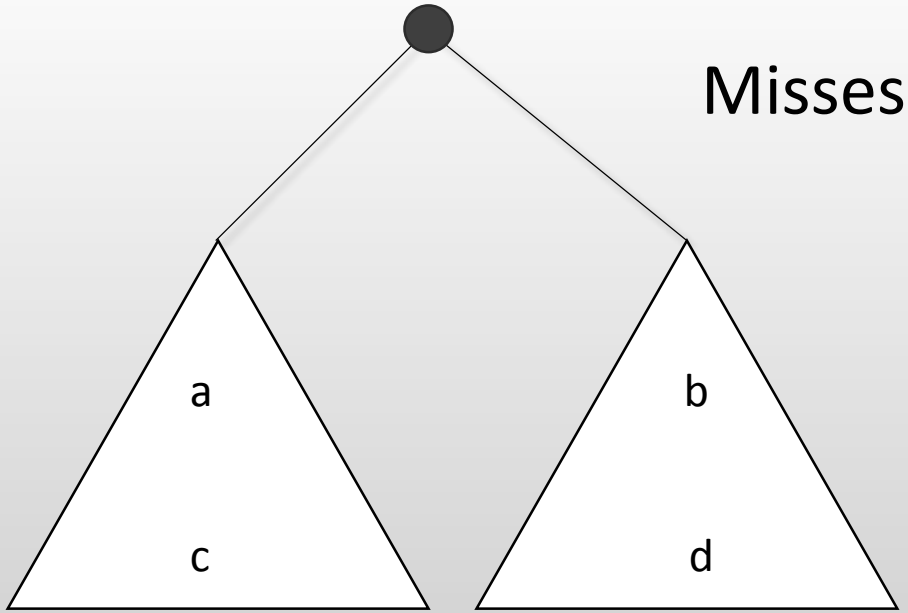
If R has a subgraph with no such cut, does it mean that we can't build a gene tree?

YES, the converse also holds.

Every cut has 2 colors → No possible rooting

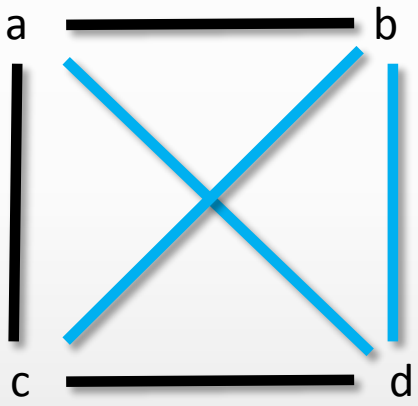


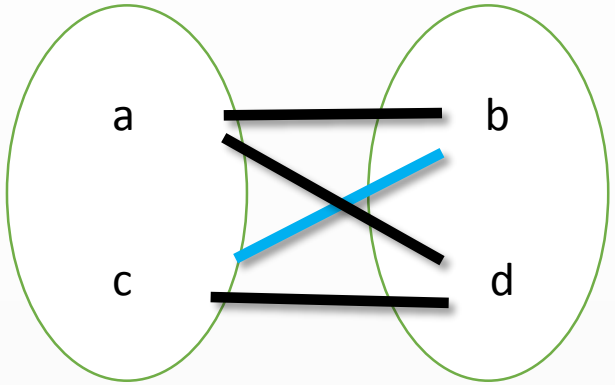


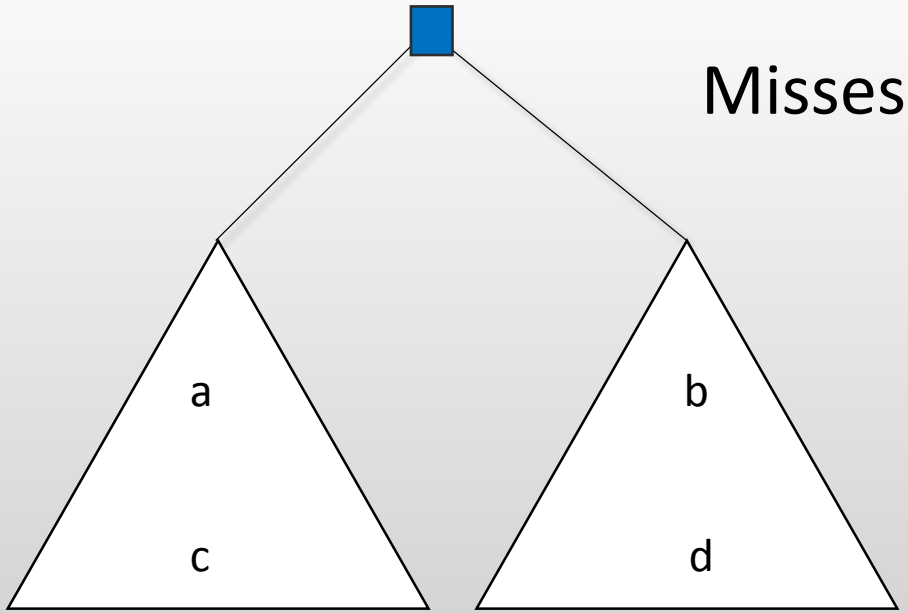


Misses the (c, b) paralogy.

Every cut has 2 colors → No possible rooting







Misses the (a, b) orthology.

Theorem:

A relation graph R is satisfiable if and only if **each subgraph has a monochromatic edge-cut**.

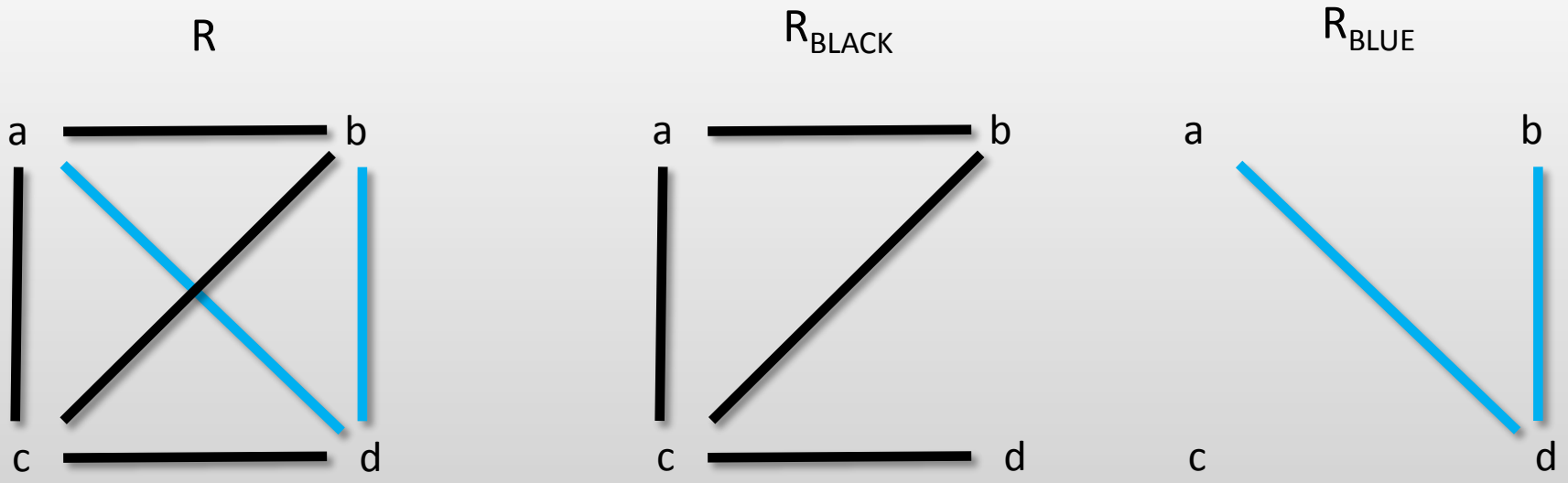
Can we test that easily (in polynomial time) ?

Theorem:

A relation graph R is satisfiable if and only if **each subgraph has a monochromatic edge-cut**.

Theorem (restated):

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



Theorem:

A relation graph R is satisfiable if and only if **each subgraph has a monochromatic edge-cut**.

Theorem (restated):

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

Theorem (again):

A relation graph R is satisfiable if and only if for each subgraph R', either R'_{BLACK} or its complement is disconnected.

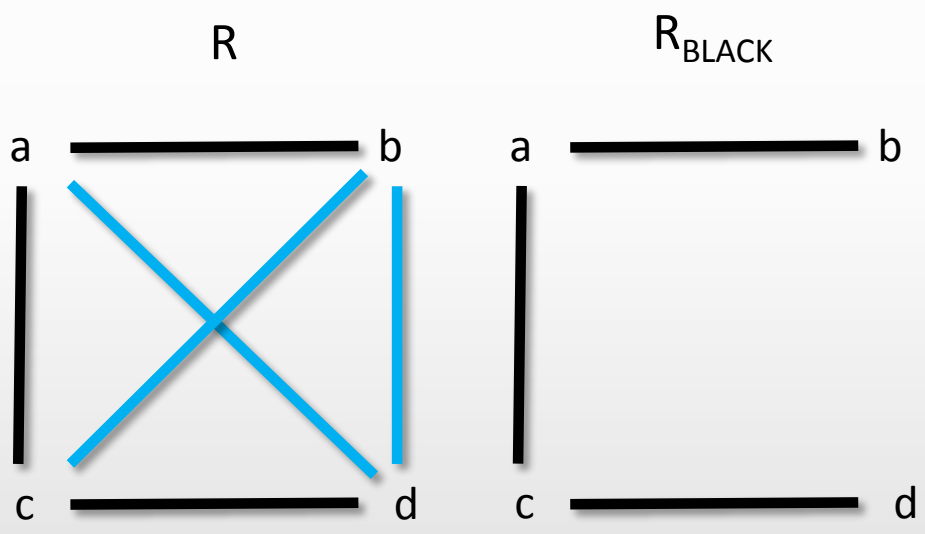
Theorem (again):

A relation graph R is satisfiable if and only if for each subgraph R', either R'_{BLACK} or its complement is disconnected.

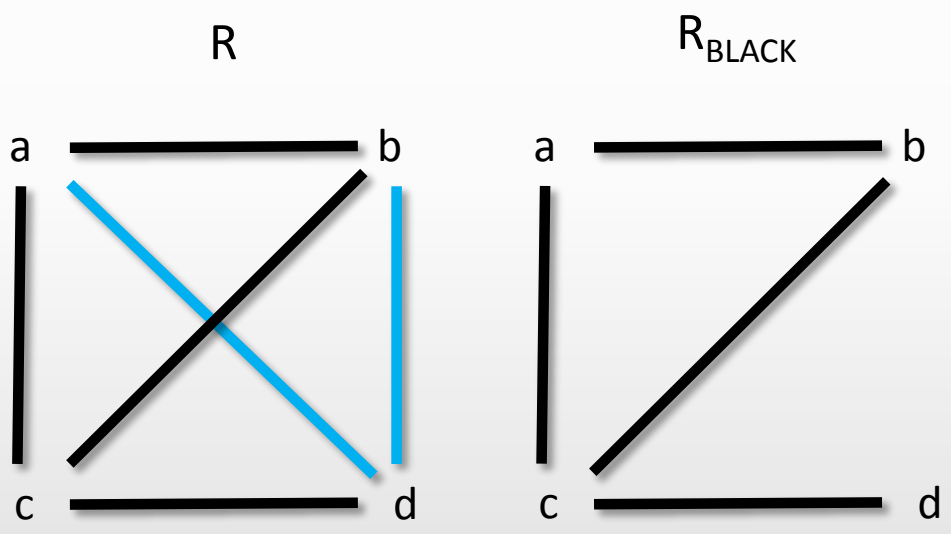
These graphs are well-known! They are called **cographs**, aka P_4-free graphs.

Theorem (finally):

A relation graph R is satisfiable if and only if R_{BLACK} is P4-free (no induced path of length 3).



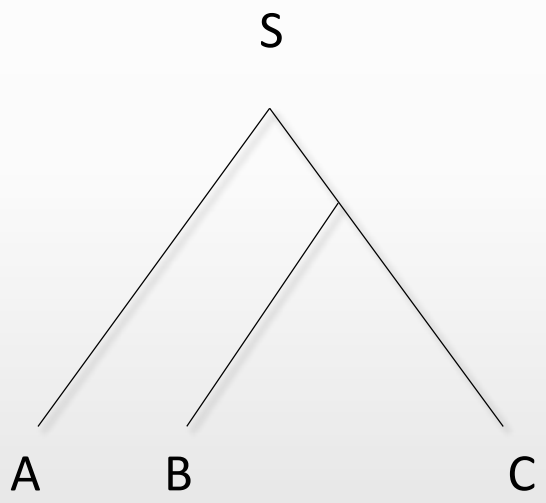
NO

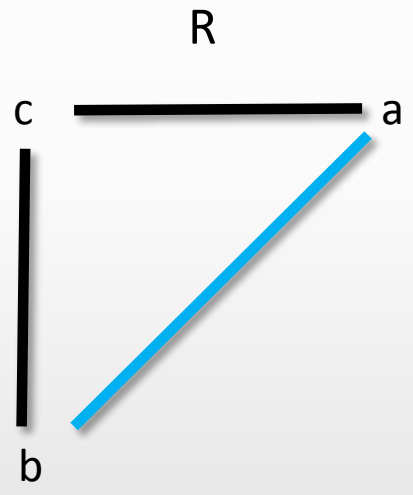


YES

S-Consistency

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

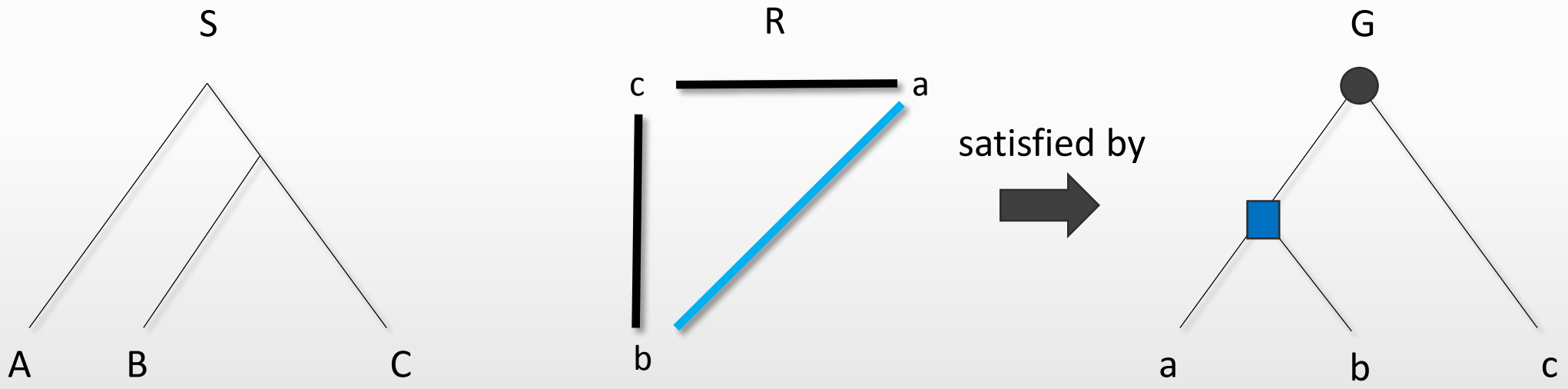




a = gene from species A
b = gene from species B
c = gene from species C

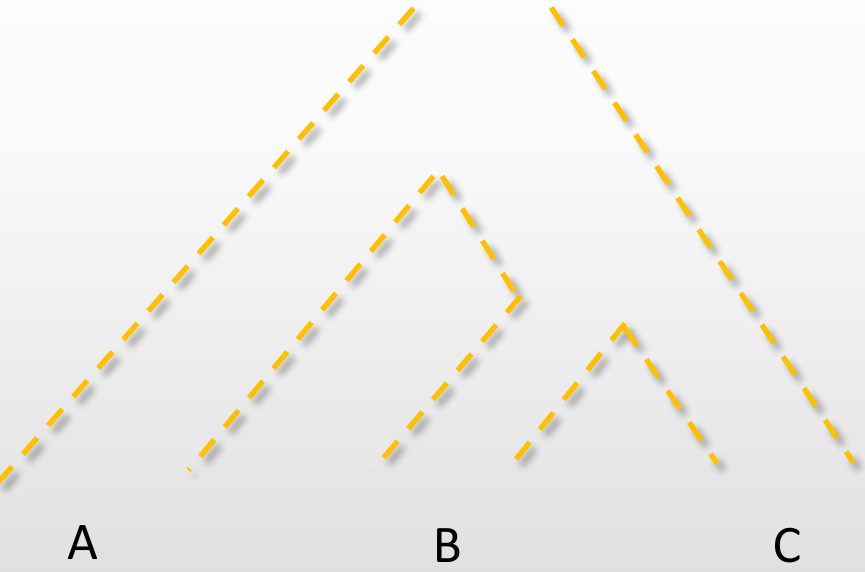
S-Consistency

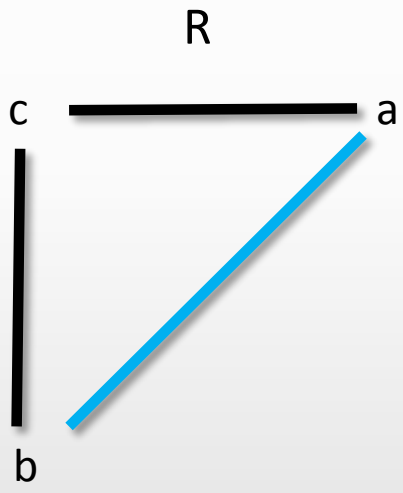
What if we want our relations to agree with a given species tree S?



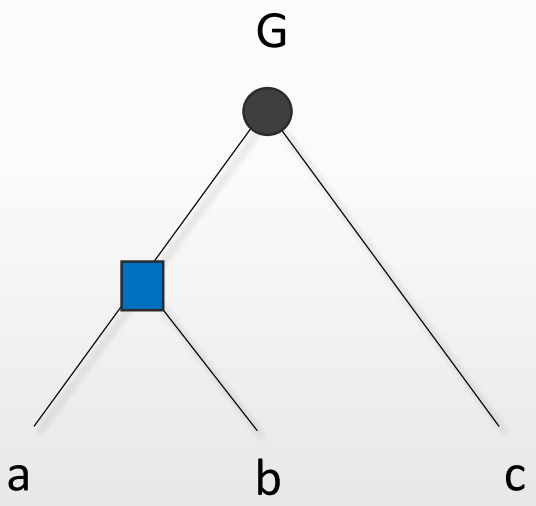
S-Consistency

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



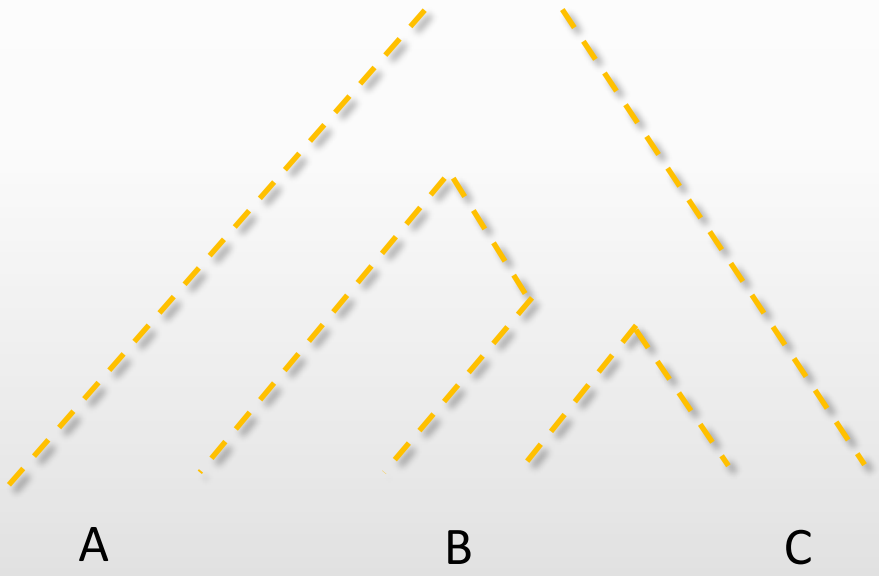


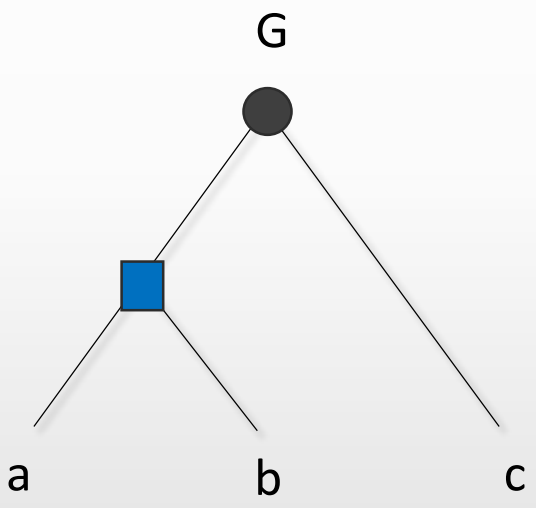
satisfied by



S-Consistency

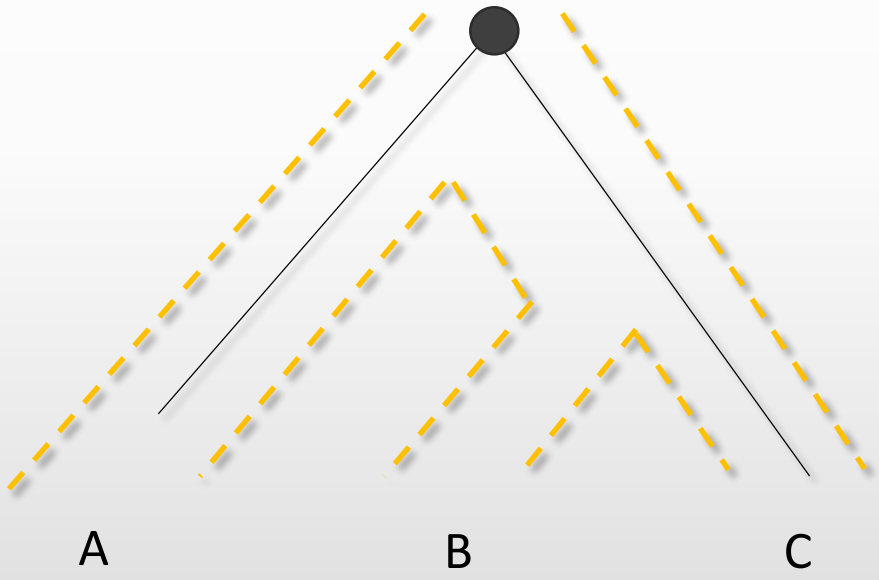
What if we want our relations to agree with a given species tree S?

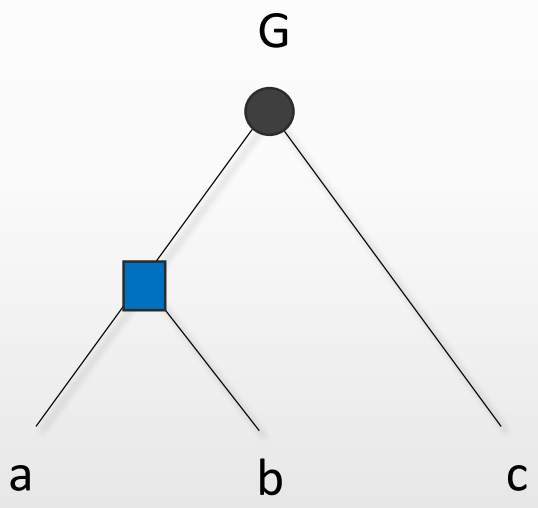




S-Consistency

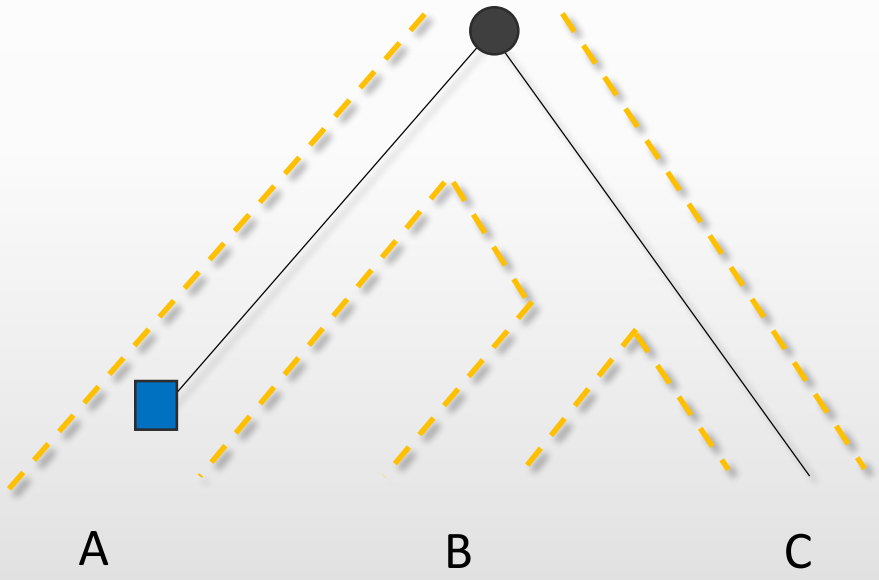
What if we want our relations to agree with a given species tree S?

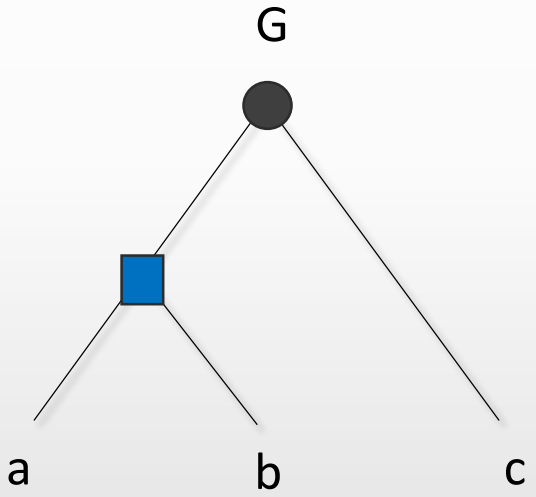




S-Consistency

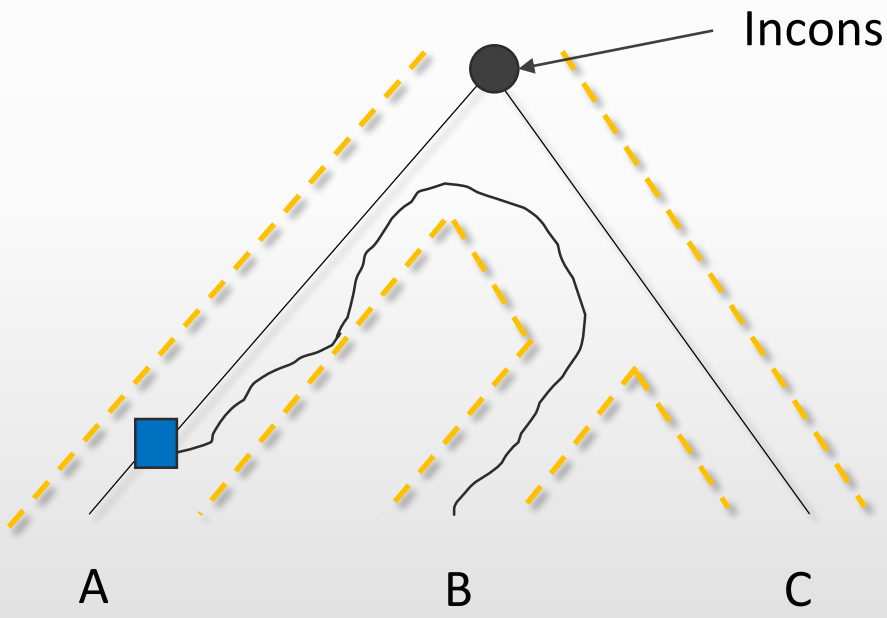
What if we want our relations to agree with a given species tree S?

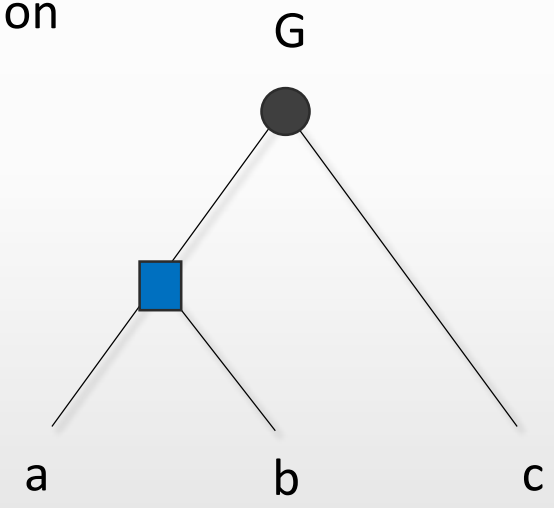




S-Consistency

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



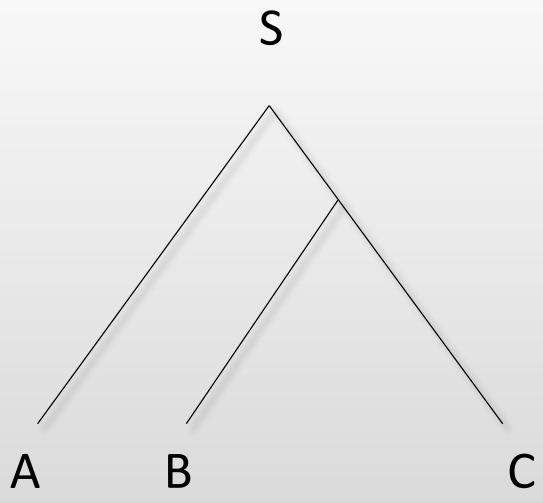


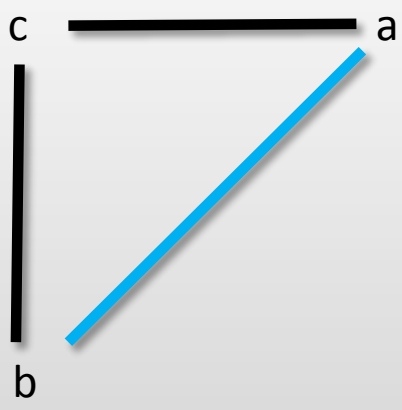
Theorem:

A relation graph R is S-Consistent if and only if R is satisfiable, **and every 3-vertex subgraph of R "agrees" with S**.

Agreement only adds a requirement on the speciations.

Only a black P_3 can possibly disagree with S.

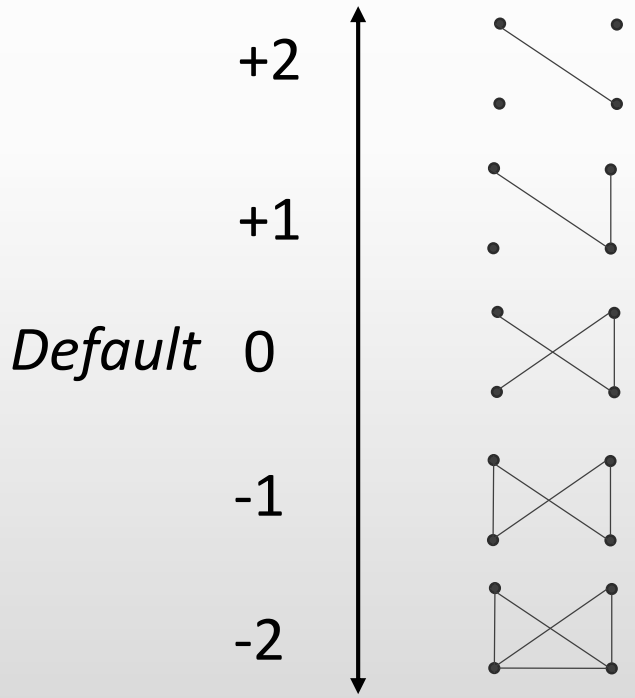




Experiments

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

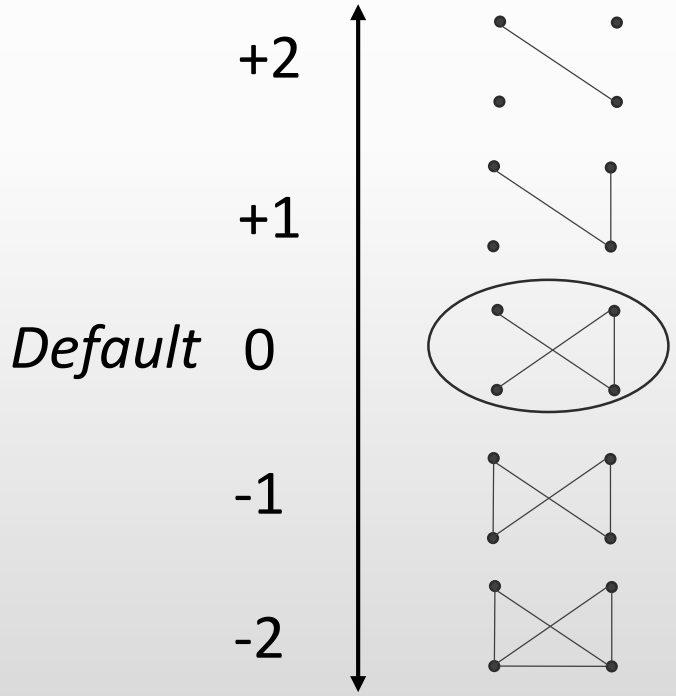
Stricter => Less orthologies



Looser => More orthologies

Experiments

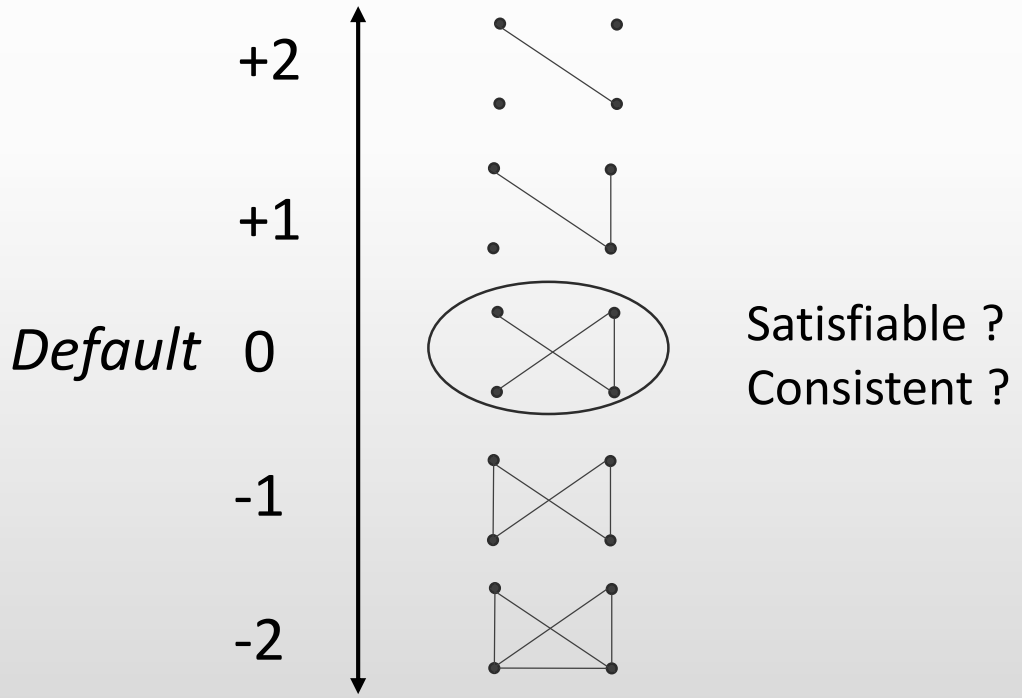
Stricter => Less orthologies



Looser => More orthologies

Experiments

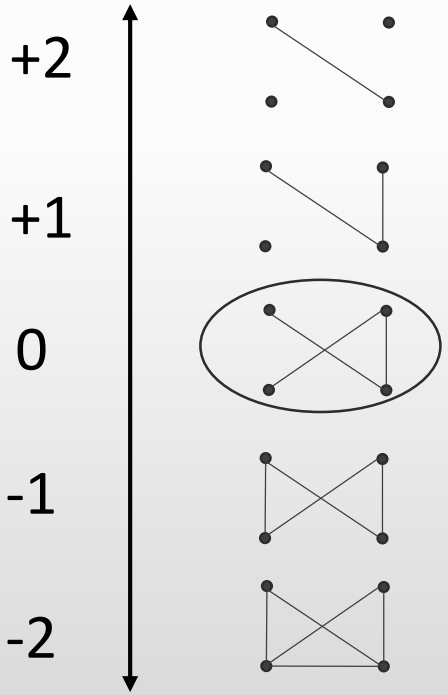
Stricter => Less orthologies



Looser => More orthologies

Experiments

Stricter => Less orthologies

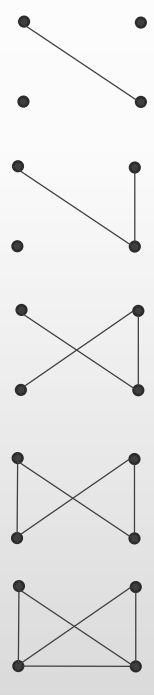


Satisfiable ? NO (~90% of families)
Consistent ? NO (~96% of families)

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%

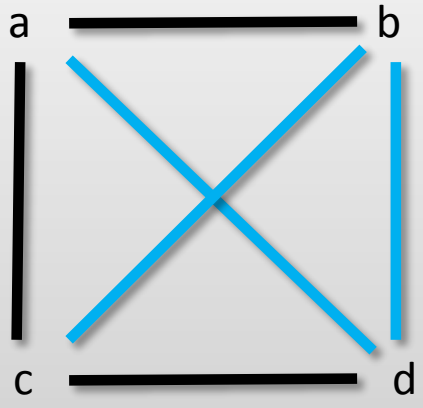
Looser => More orthologies

Gene relation correction

Gene relation correction

Make R satisfiable by changing a minimum number of relations.

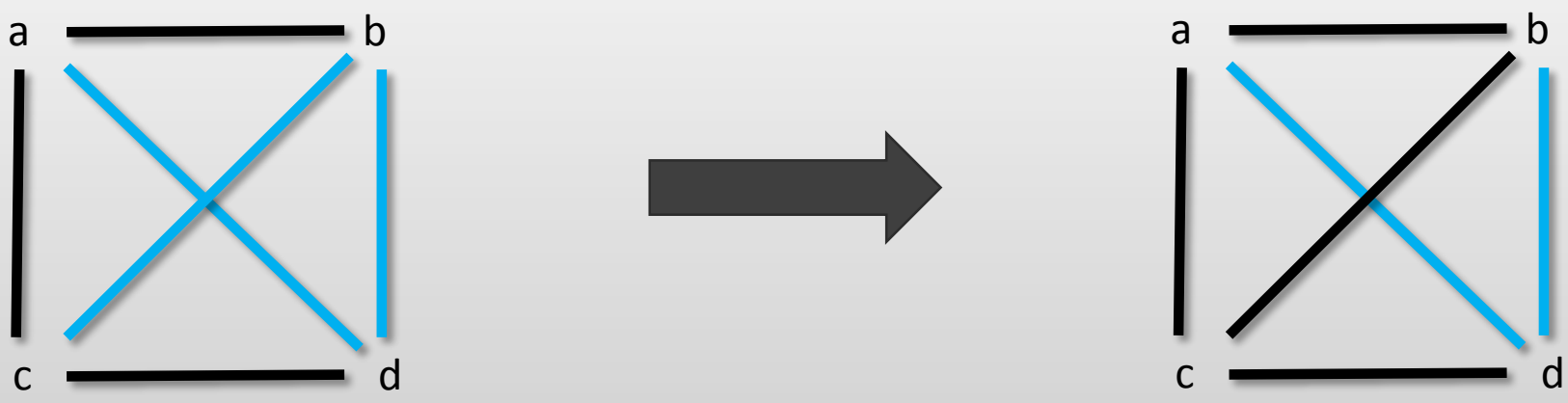
That is, change as few edge colors as possible to make R_{BLACK} P_4-free



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

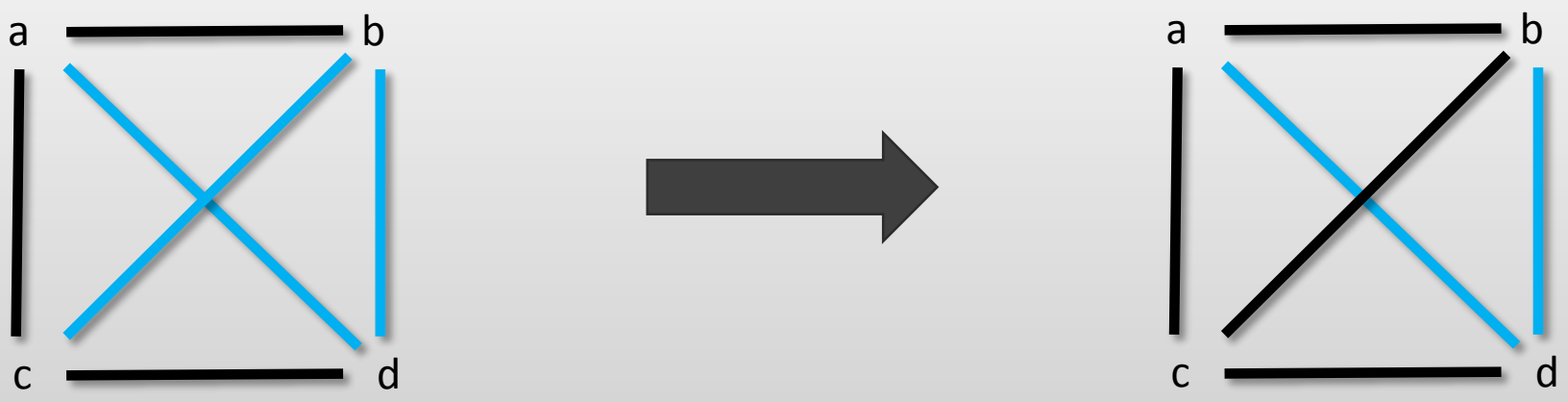


Gene relation correction

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

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.

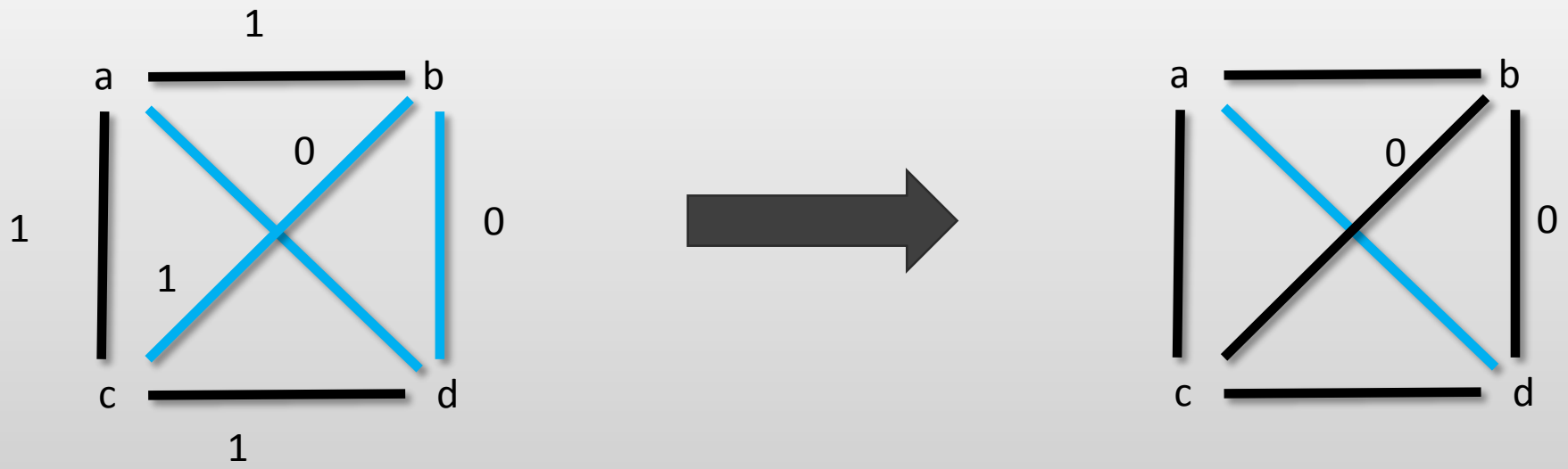
Weighted gene relation correction

Something we can handle:

If edges all have weights of 0 or 1

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

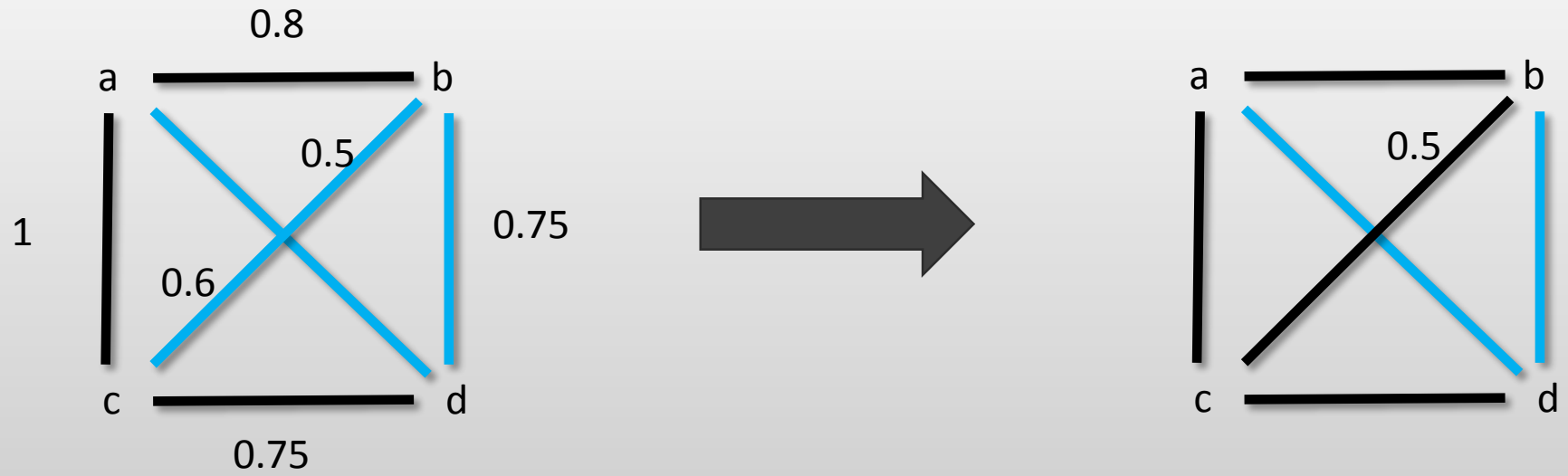
We can tell in polynomial time if there is an edge editing of weight 0.



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*.



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.

Self-consistency: we don't know the species tree S, but we want the relations to be consistent with some species tree.

HGT, ILS, etc. : how can we handle other events such as horizontal gene trasnfer or incomplete lineage sorting? What are their impact on relation graphs?

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?