THÉORIE DES GRAPHES ET ÉVOLUTION: UNE CARACTÉRISATION DES GRAPHES D'ORTHOLOGIE

Manuel Lafond

UNIVERSITÉ DE SHERBROOKE

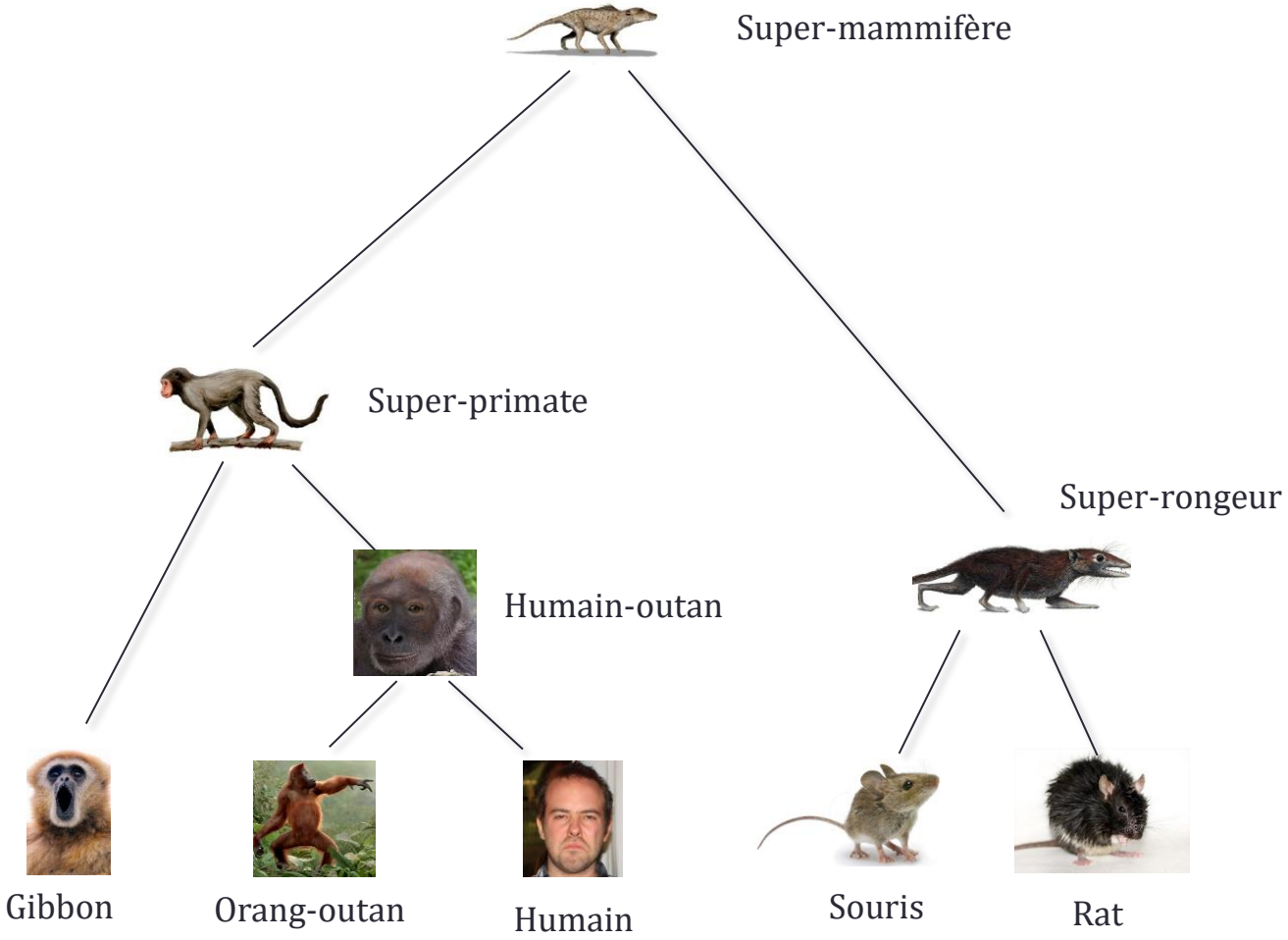
Contexte

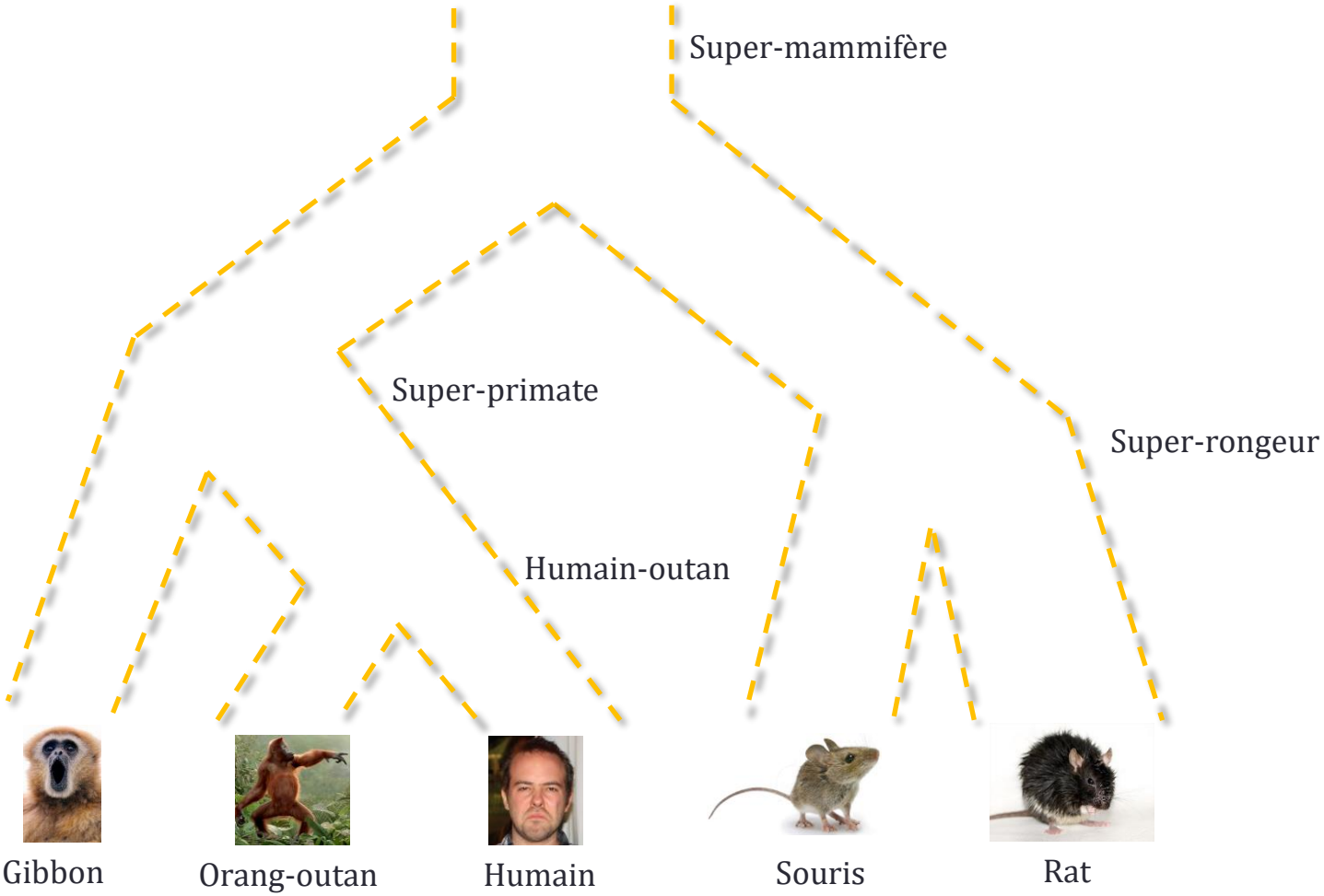
- Évolution des gènes et des espèces.
- Gènes orthologues et graphes d'orthologie.

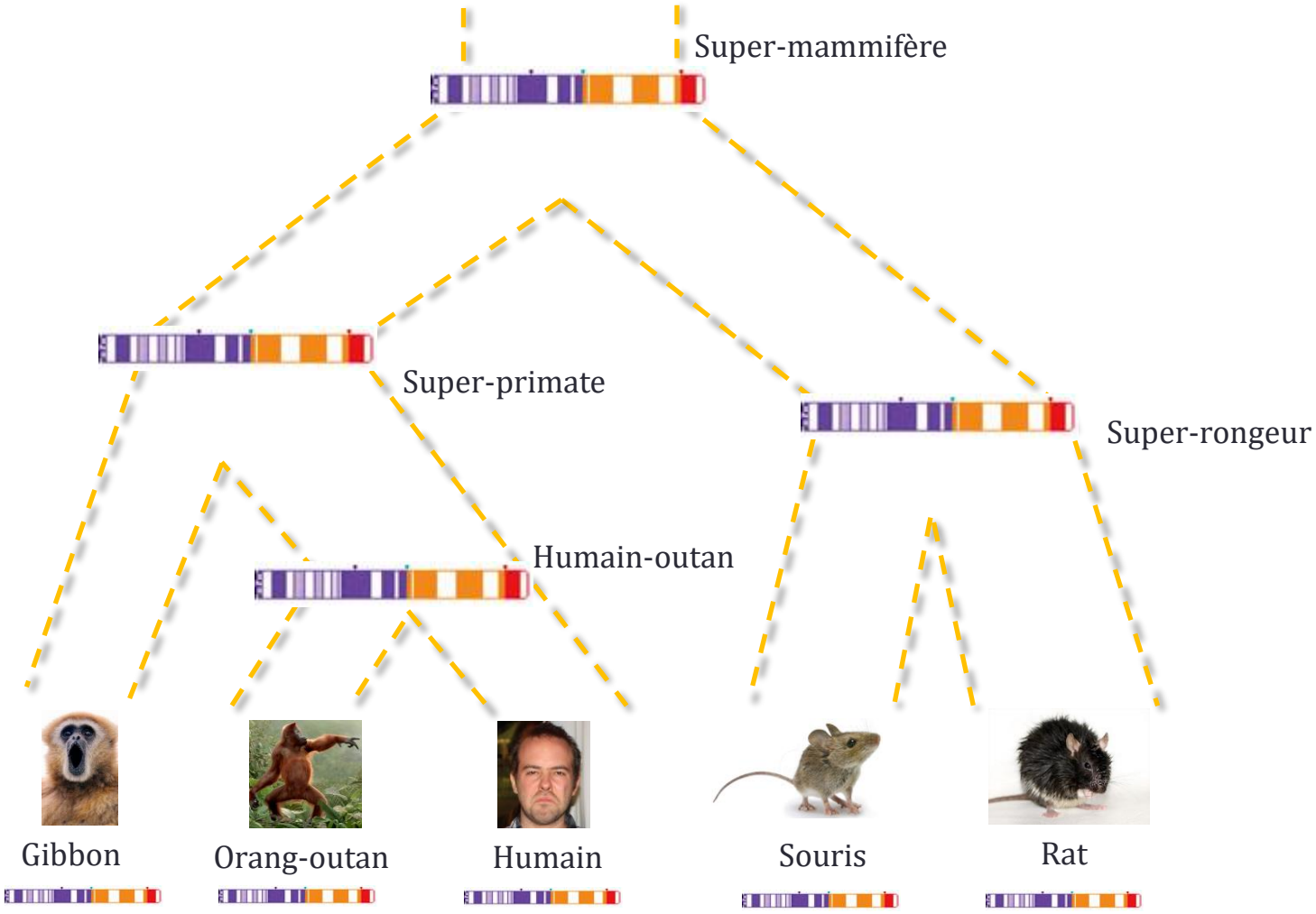
Résultat

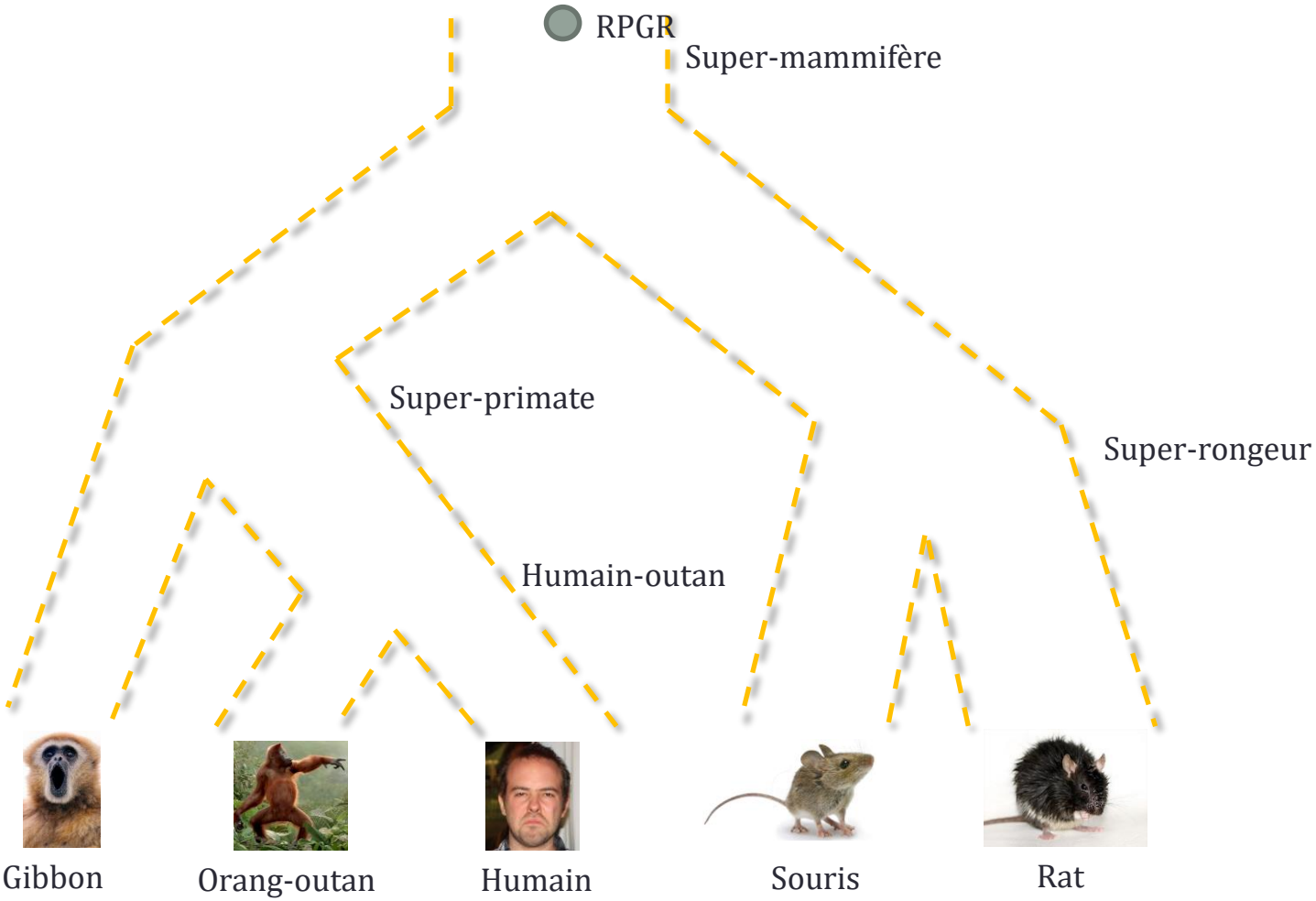
- Les graphes d'orthologie sont précisément les graphes sans P4.

Évolution des gènes et espèces

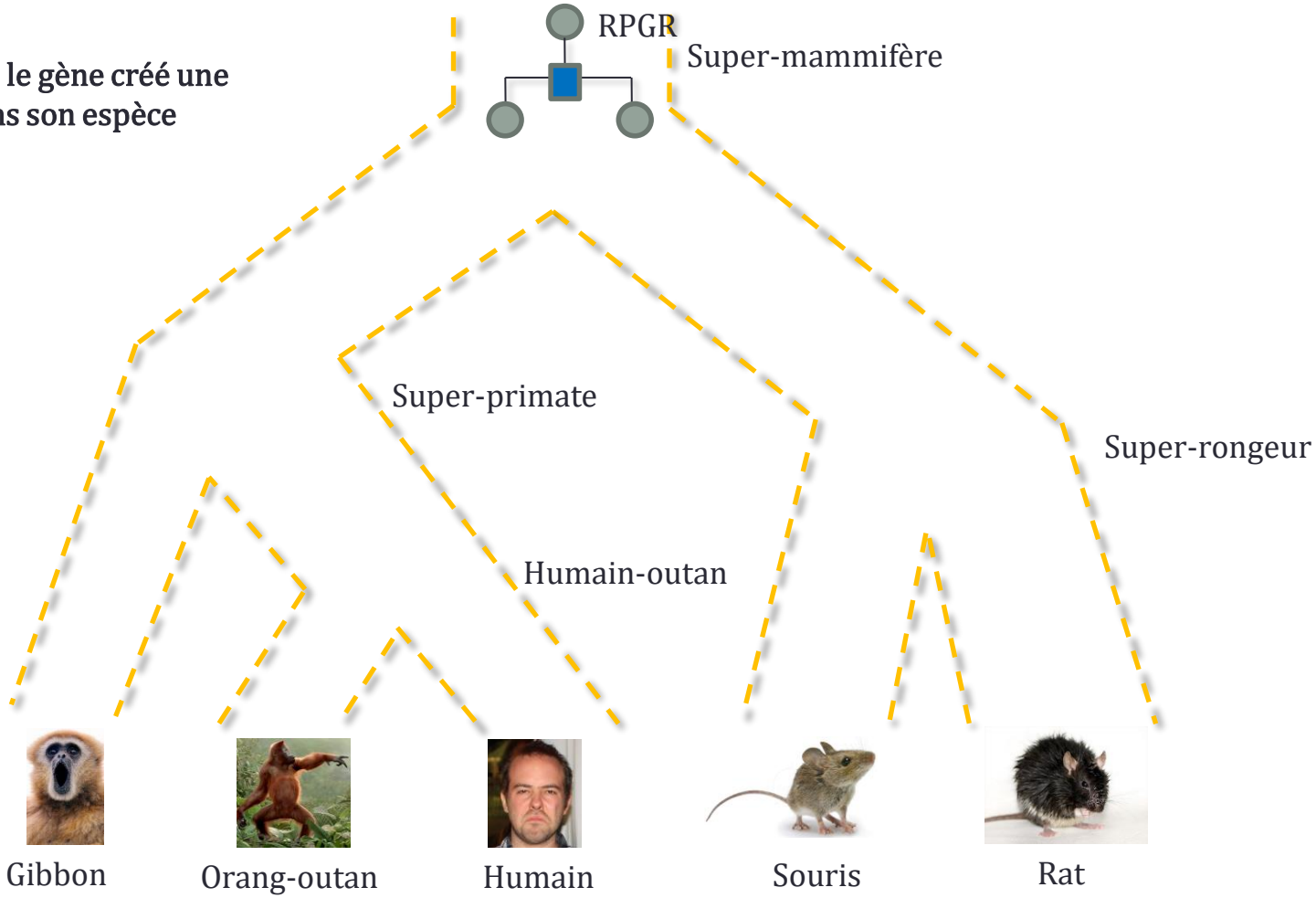



Super-mammifère
Super-primate
Super-rongeur
Humain-outan
Gibbon
Orang-outan
Humain
Souris
Rat

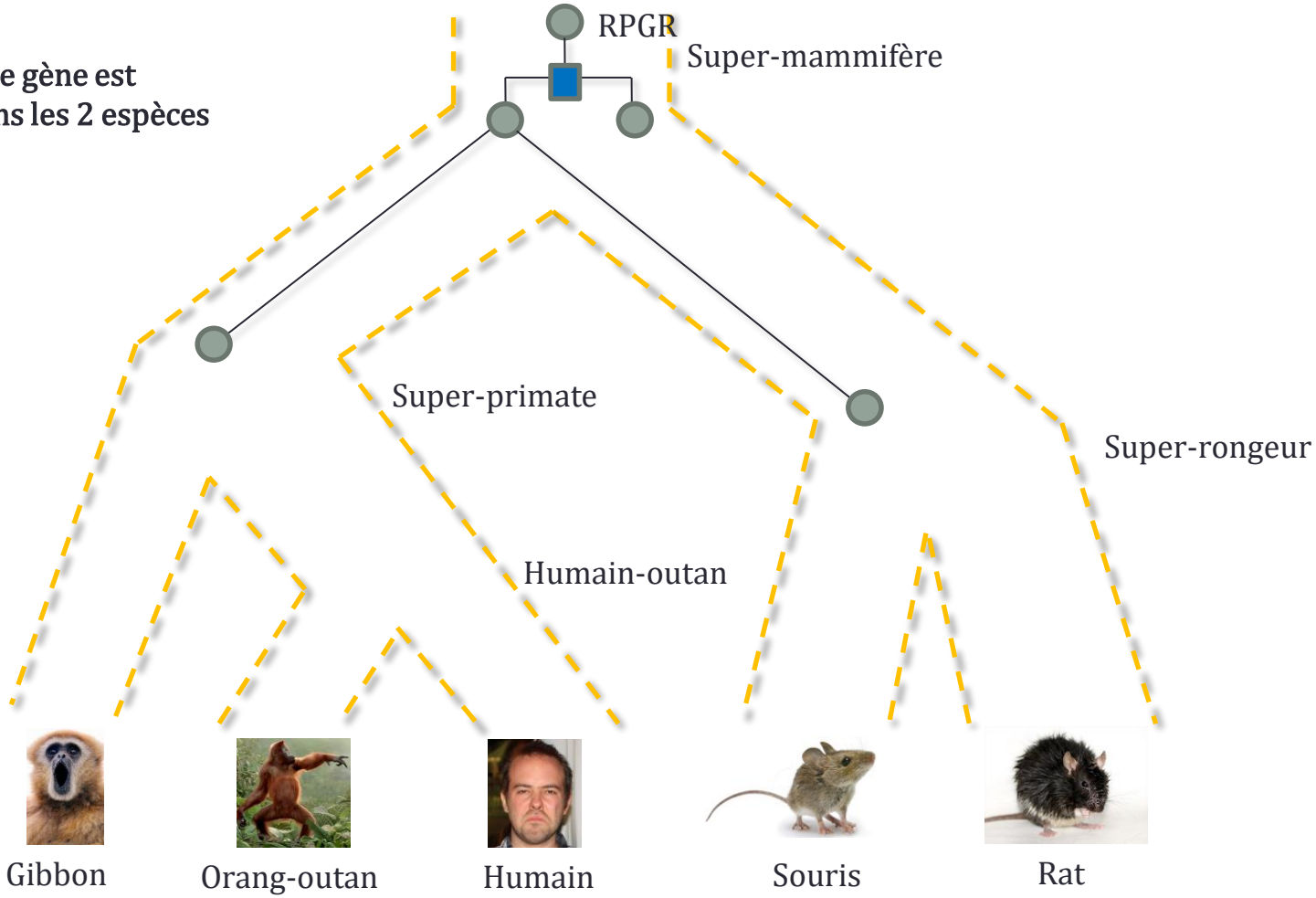

Super-mammifère
Super-primate
Super-rongeur
Humain-outan
Gibbon
Orang-outan
Humain
Souris
Rat

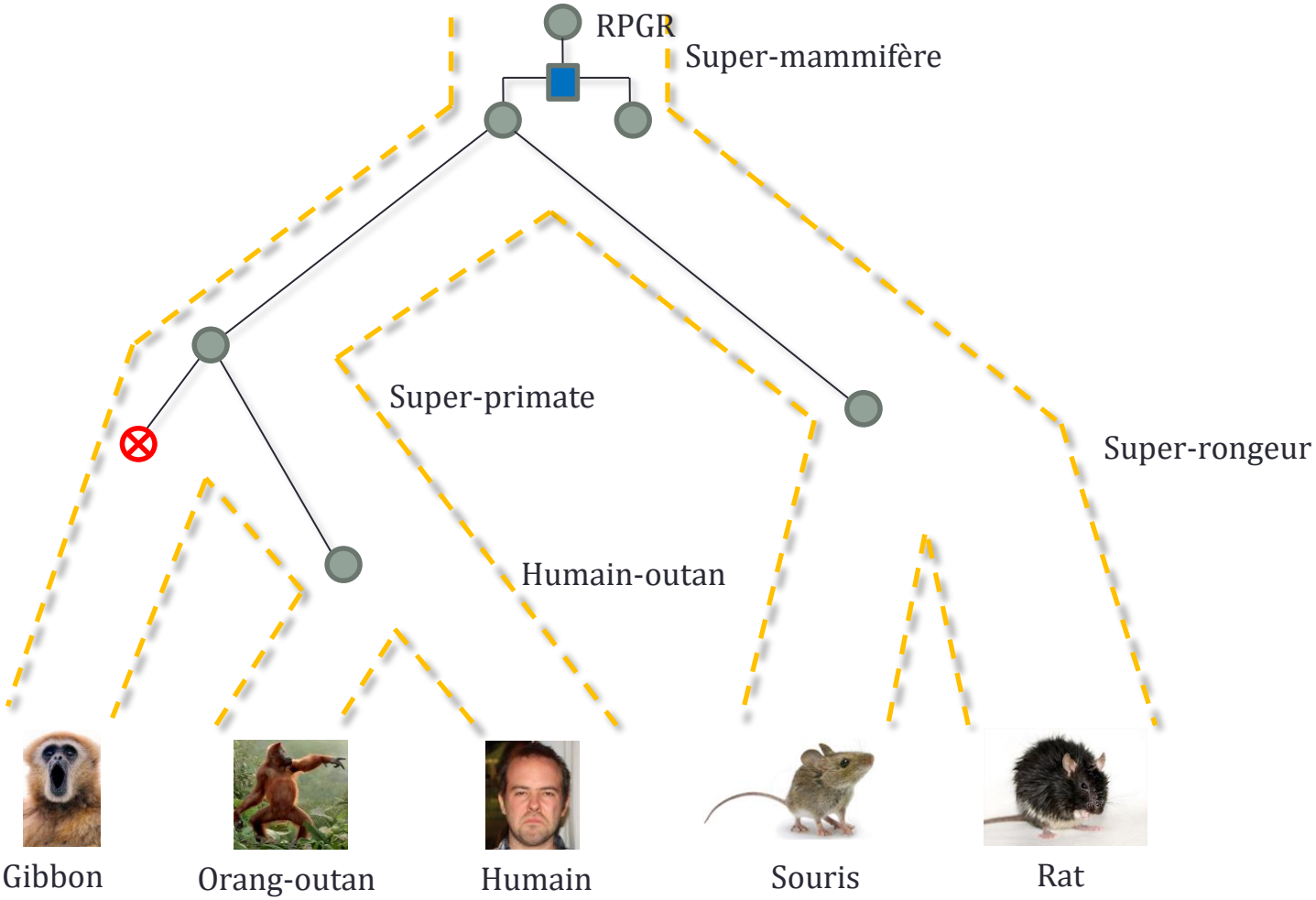

RPGR
Super-mammifère
Super-primate
Super-rongeur
Humain-outan
Gibbon
Orang-outan
Humain
Souris
Rat

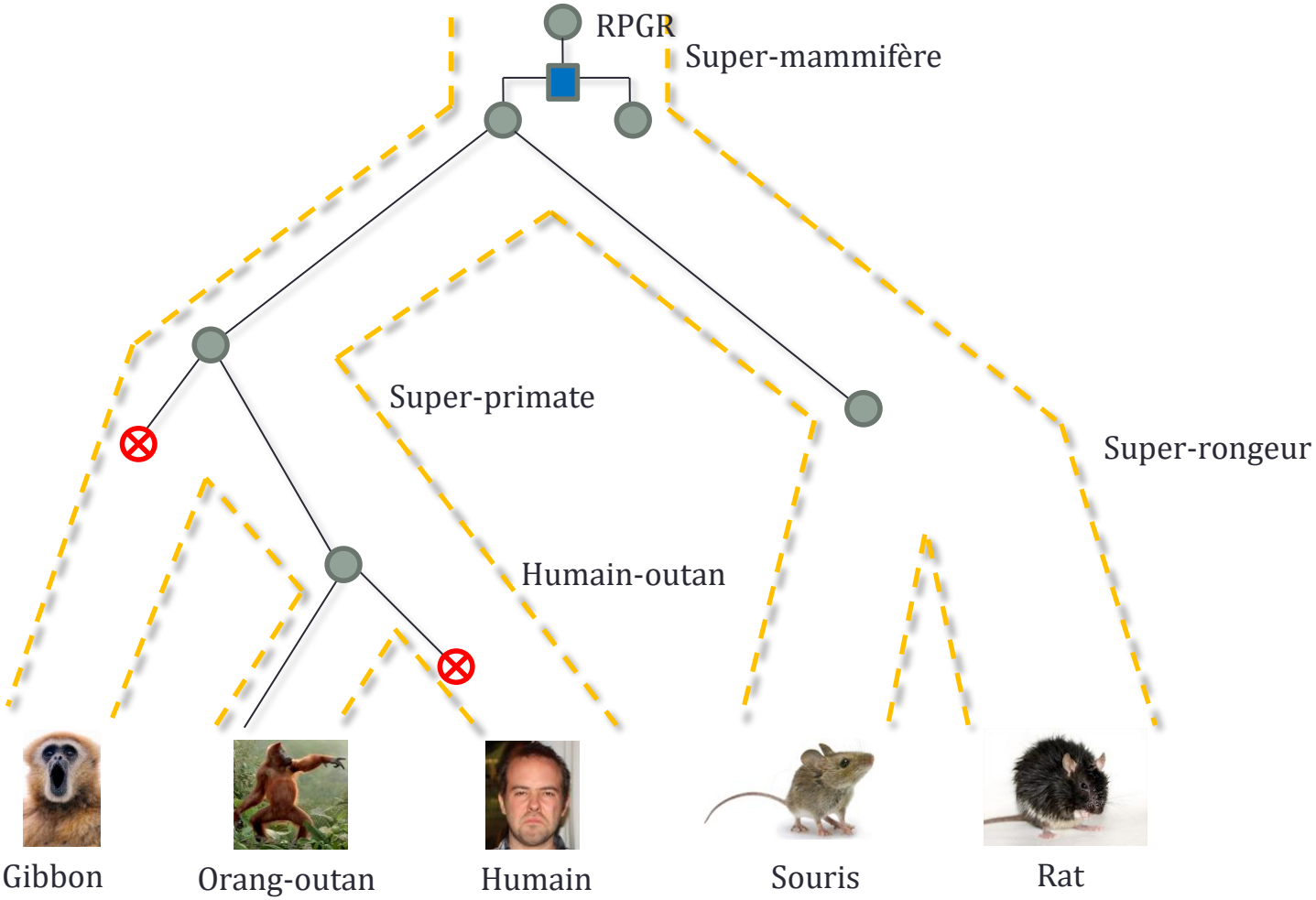
Duplication = le gène créé une 2ème copie dans son espèce

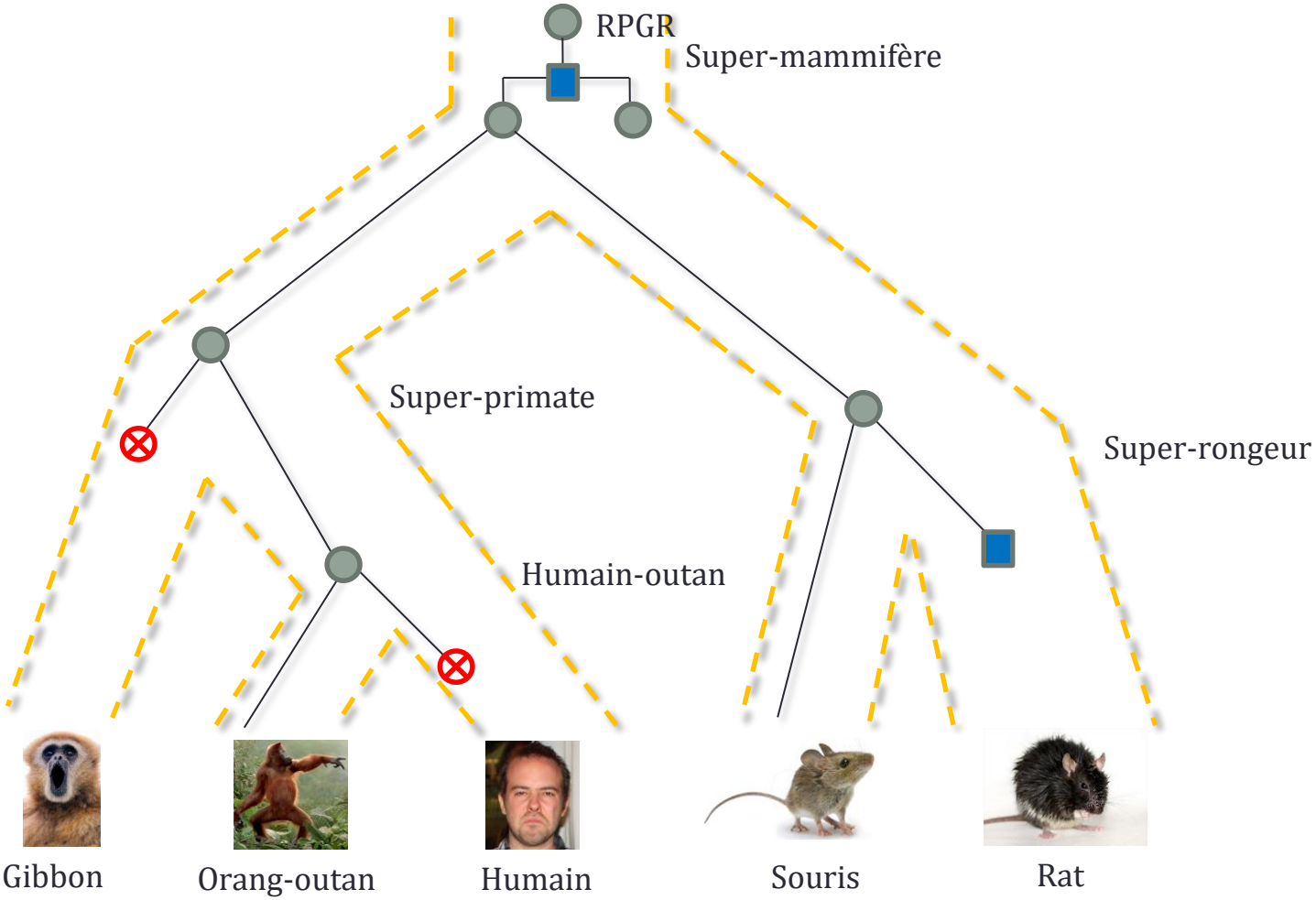


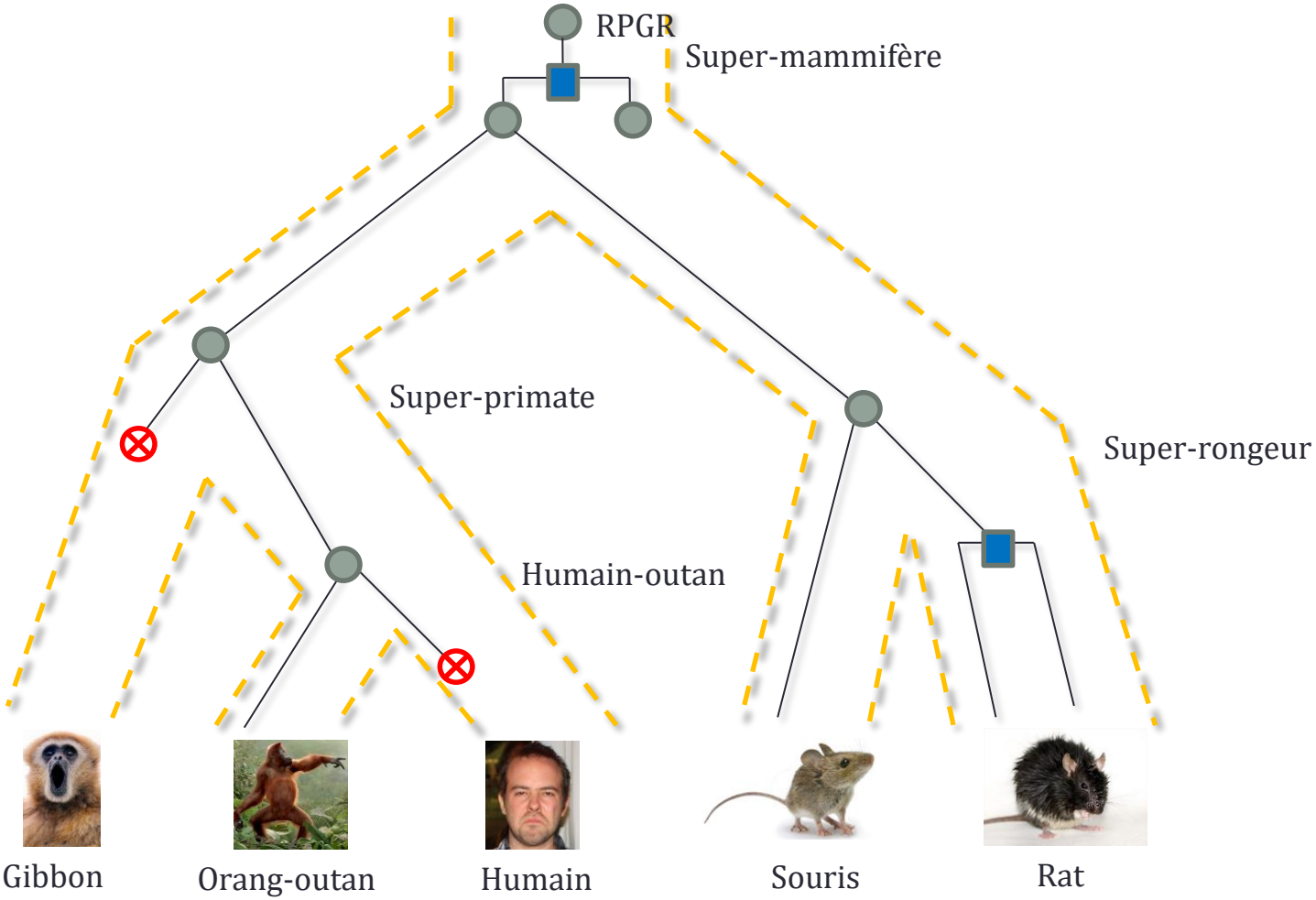
Spéciation = le gène est « envoyé » dans les 2 espèces descendantes

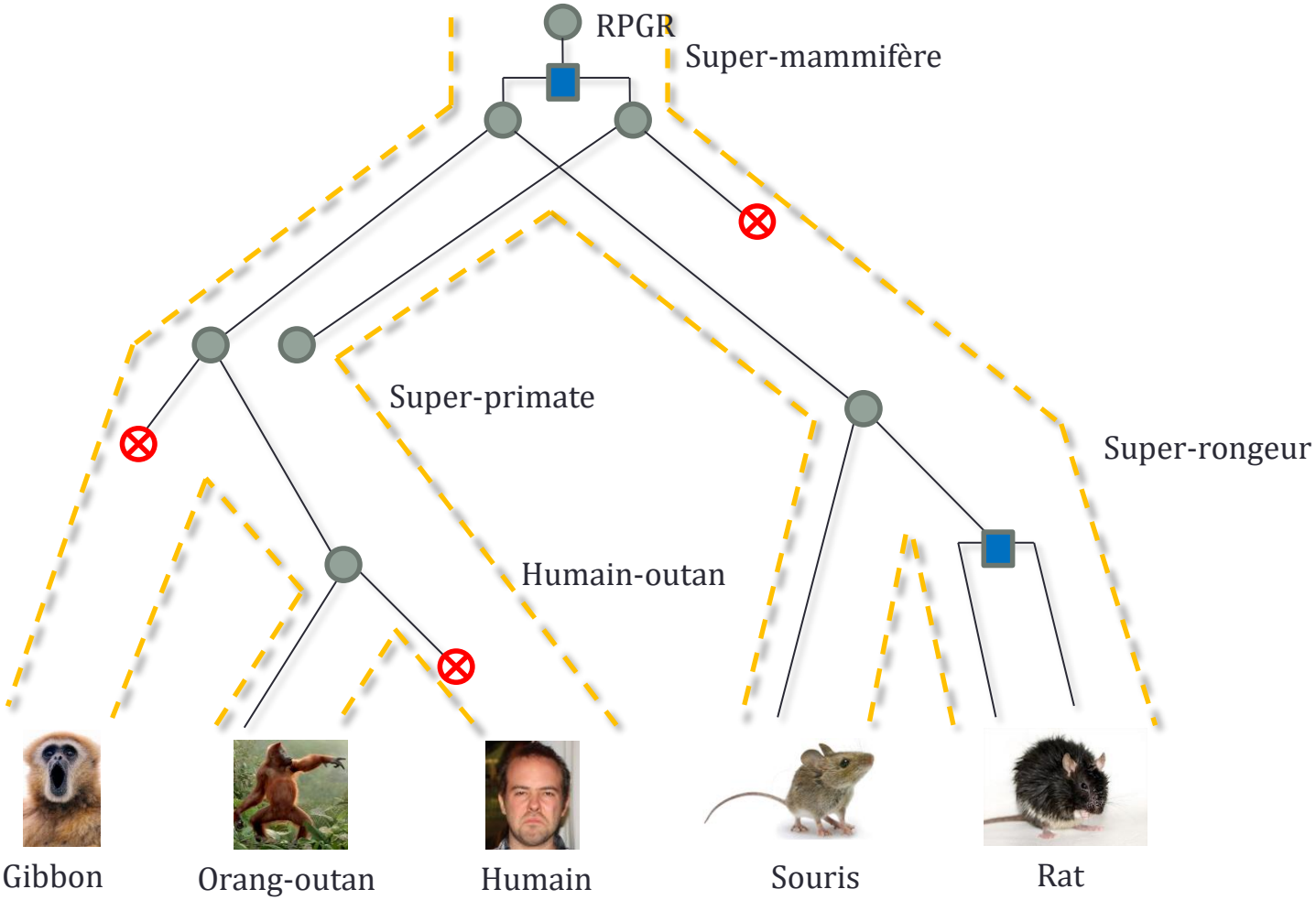


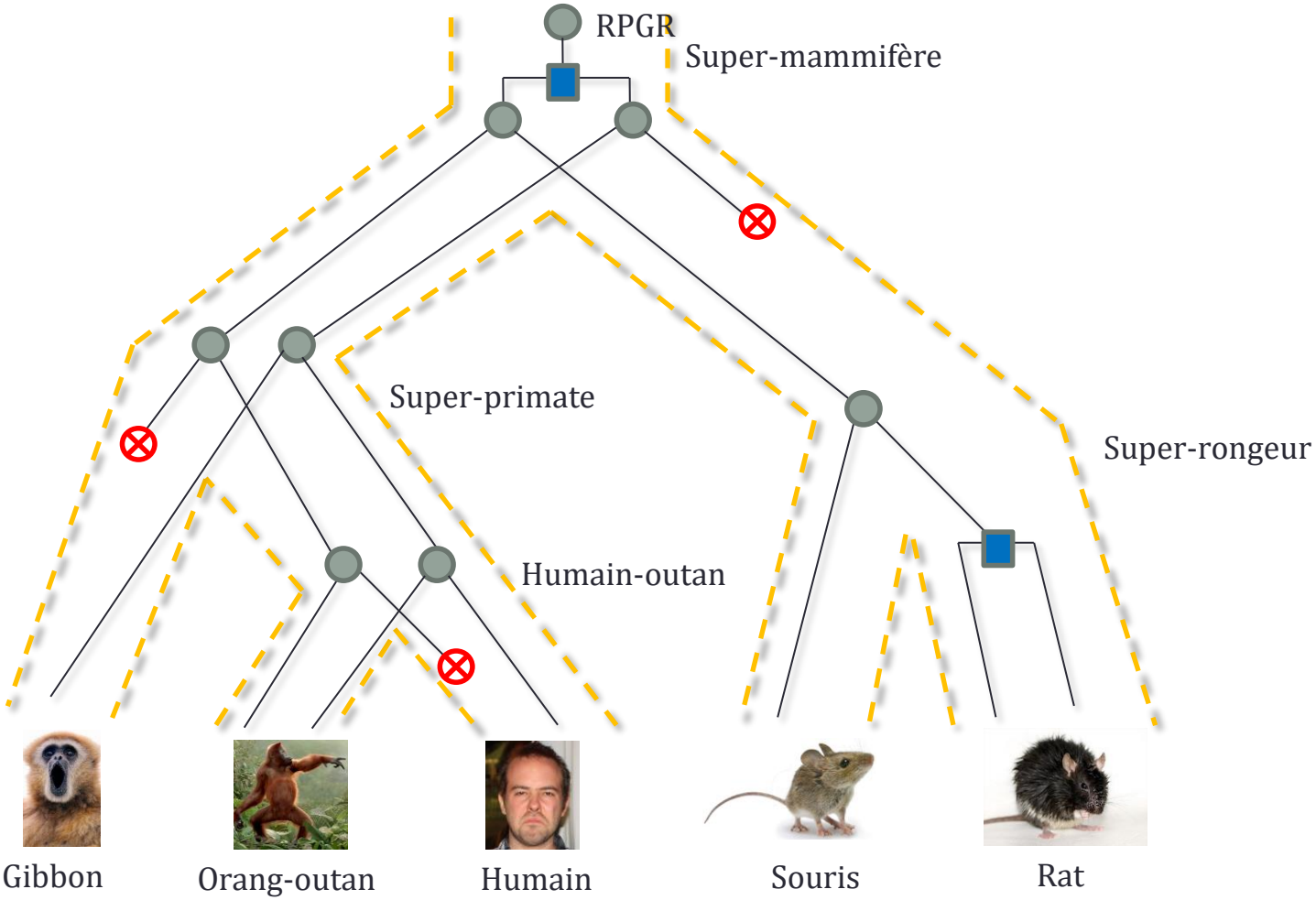

RPGR
Super-mammifère
Super-primate
Super-rongeur
Humain-outan
Gibbon
Orang-outan
Humain
Souris
Rat

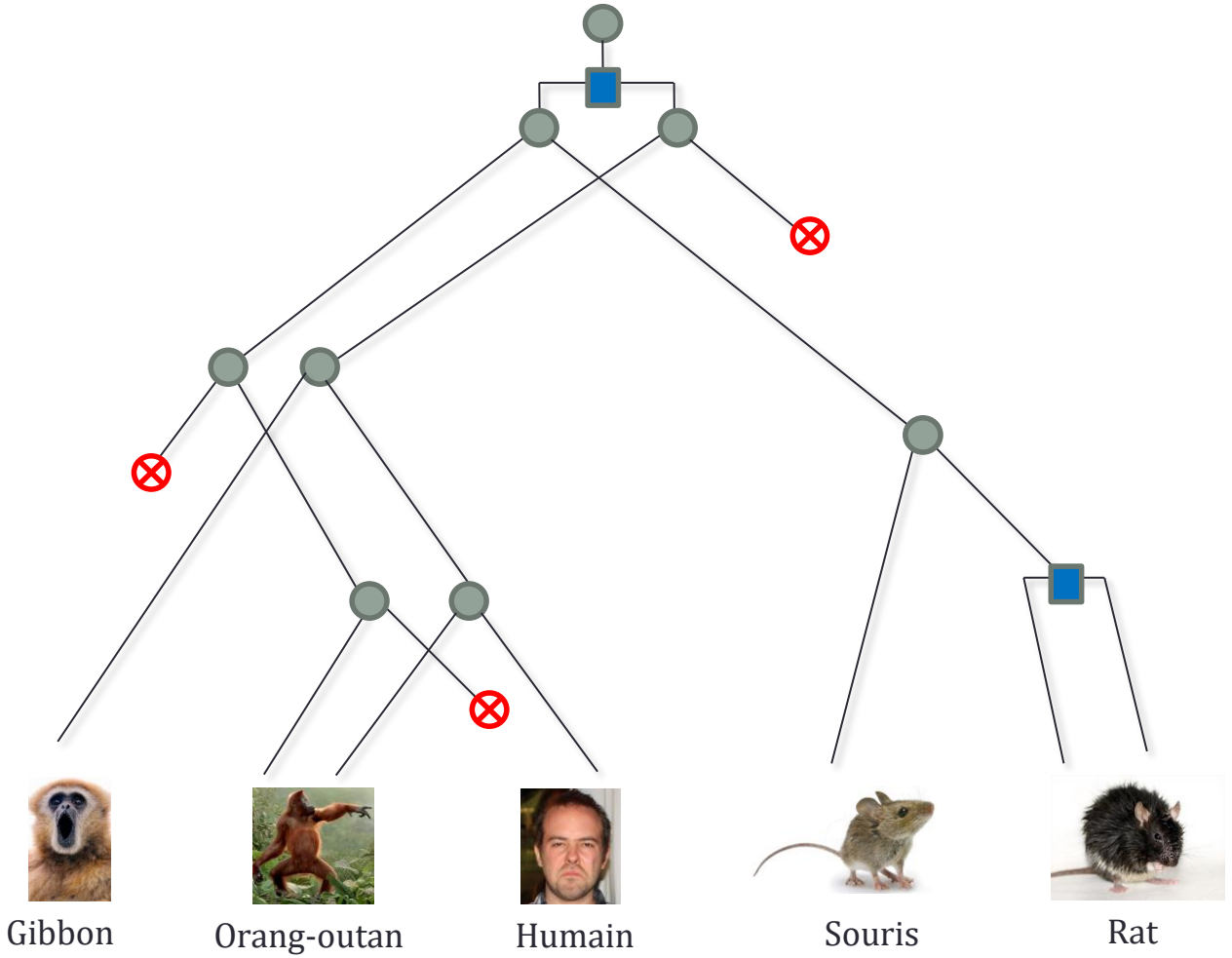

RPGR
Super-mammifère
Super-primate
Super-rongeur
Humain-outan
Gibbon
Orang-outan
Humain
Souris
Rat


RPGR
Super-mammifère
Super-primate
Super-rongeur
Humain-outan
Gibbon
Orang-outan
Humain
Souris
Rat

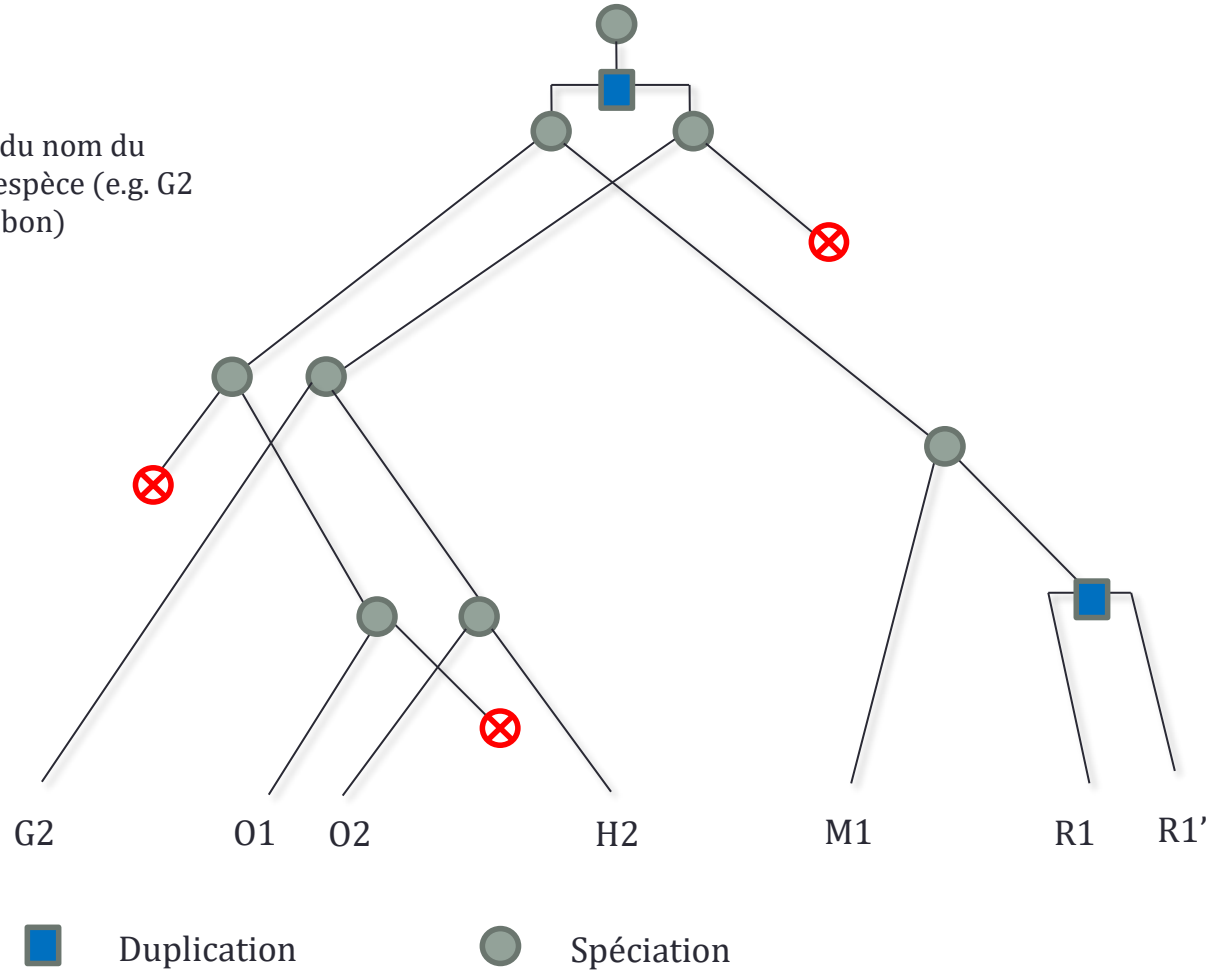

RPGR
Super-mammifère
Super-primate
Super-rongeur
Humain-outan
Gibbon
Orang-outan
Humain
Souris
Rat

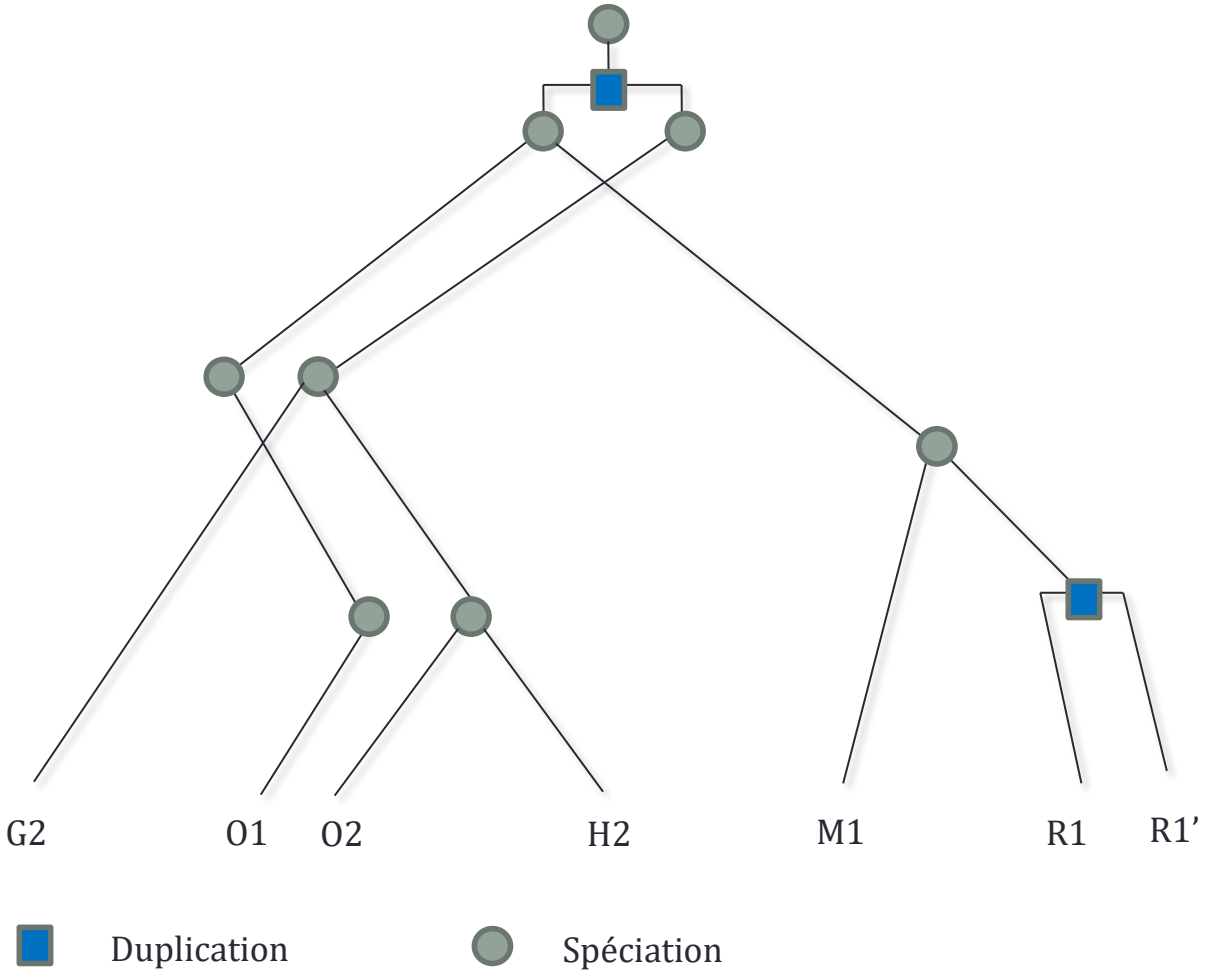

RPGR
Super-mammifère
Super-primate
Super-rongeur
Humain-outan
Gibbon
Orang-outan
Humain
Souris
Rat

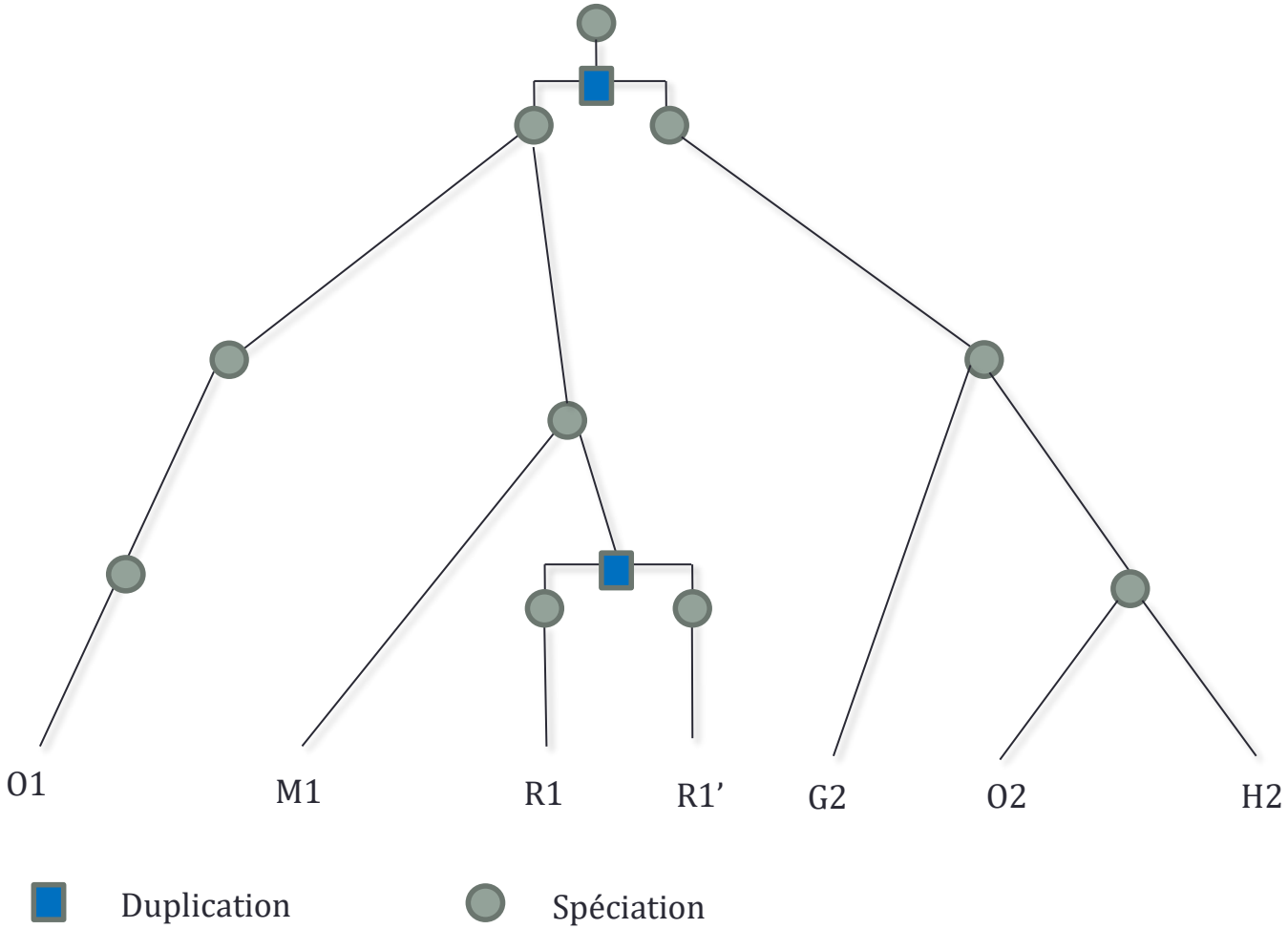

RPGR
Super-mammifère
Super-primate
Super-rongeur
Humain-outan
Gibbon
Orang-outan
Humain
Souris
Rat

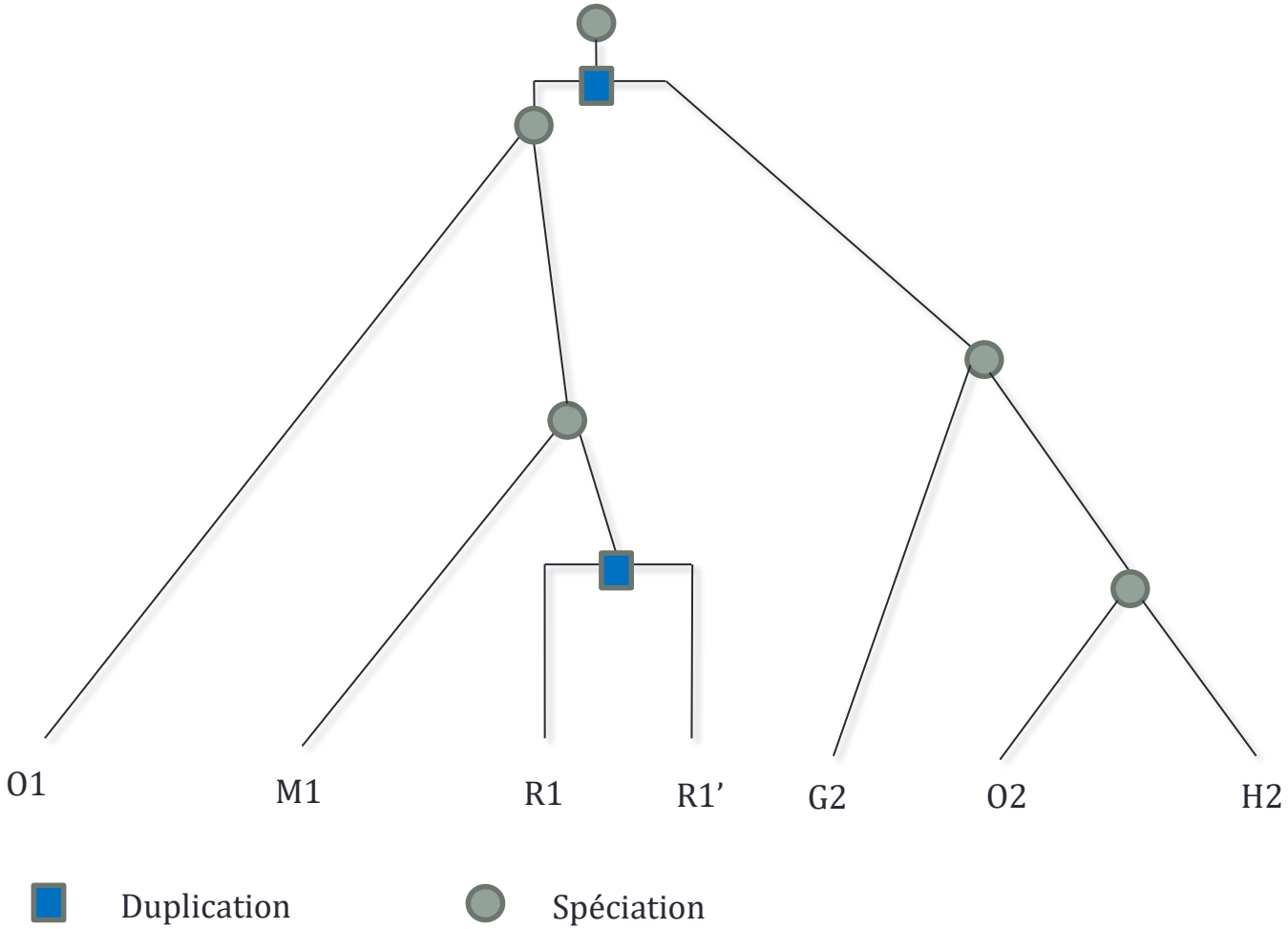

Gibbon
Orang-outan
Humain
Souris
Rat

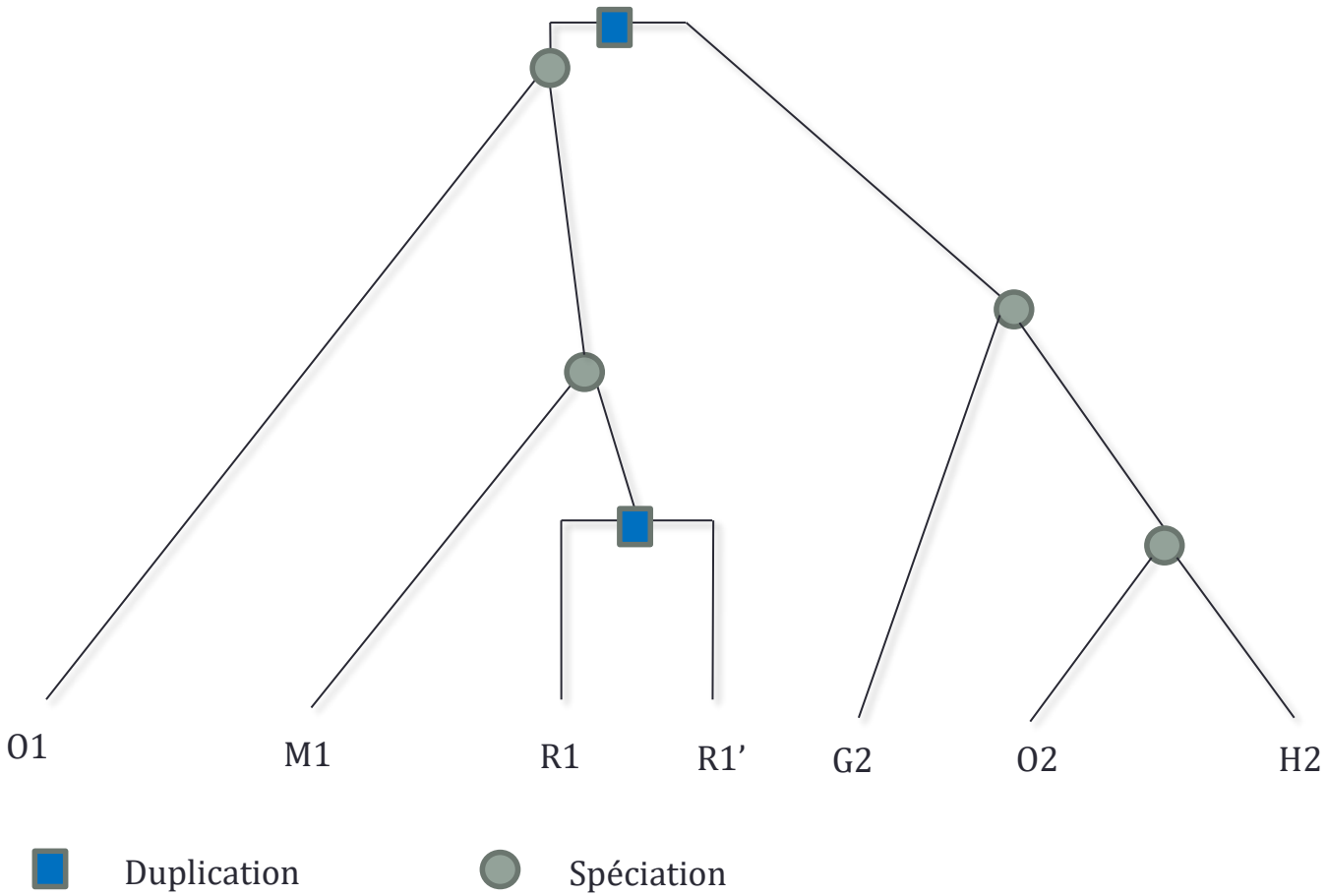
Note: la 1re lettre du nom du gène dénote son espèce (e.g. G2 appartient au **G**ibbon)

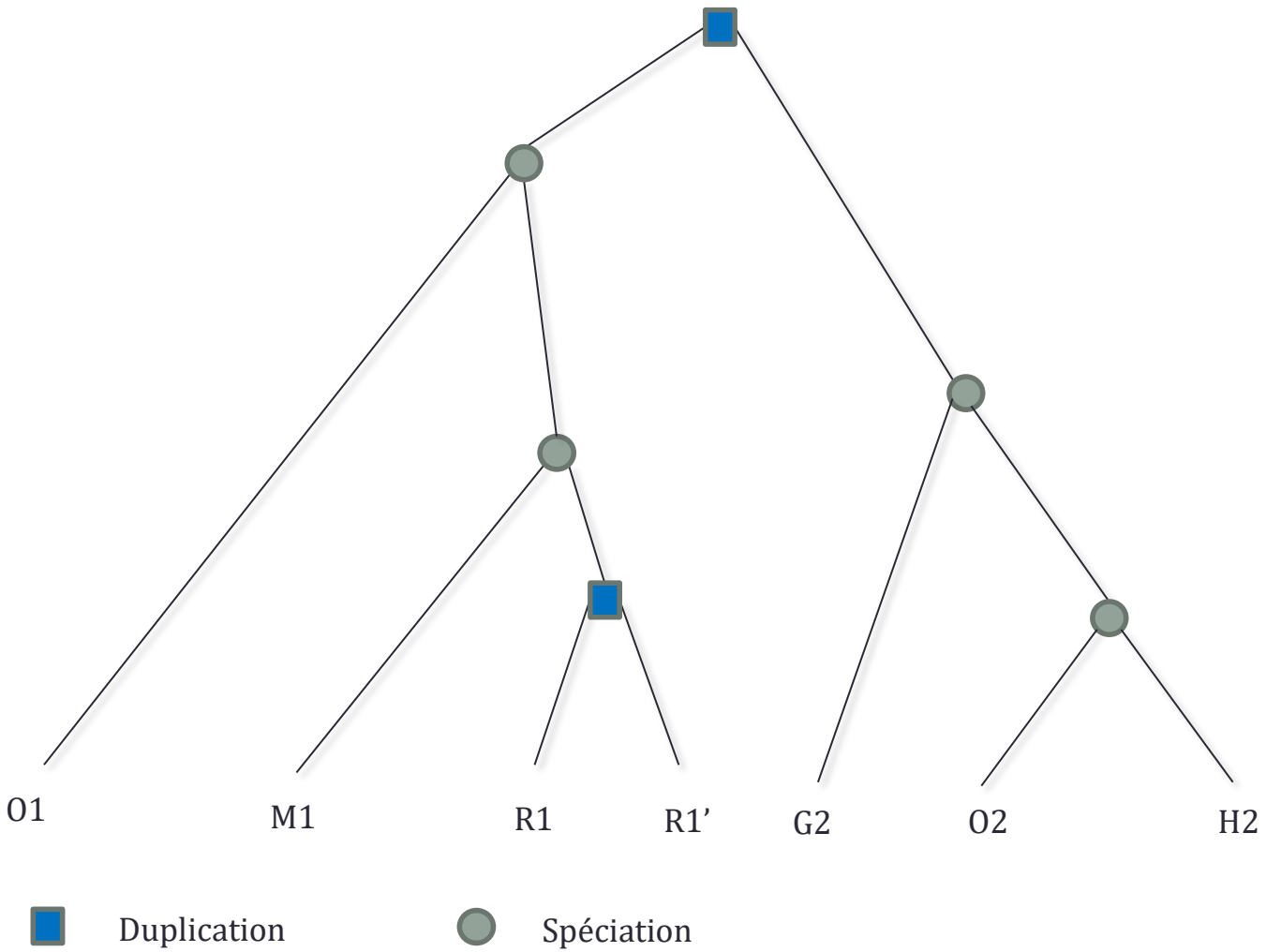



G2
O1
O2
H2
M1
R1
R1’
Duplication
Spéciation


O1
M1
R1
R1'
G2
O2
H2
Duplication
Spéciation


O1
M1
R1
R1'
G2
O2
H2
Duplication
Spéciation


O1
M1
R1
R1'
G2
O2
H2
Duplication
Spéciation


O1
M1
R1
R1'
G2
O2
H2
Duplication
Spéciation

Orthologues et paralogues

L'arbre de gènes définit une relation binaire.

Deux gènes sont:

Orthologues si leur **ancêtre commun le plus récent** a subit une **spéciation**.

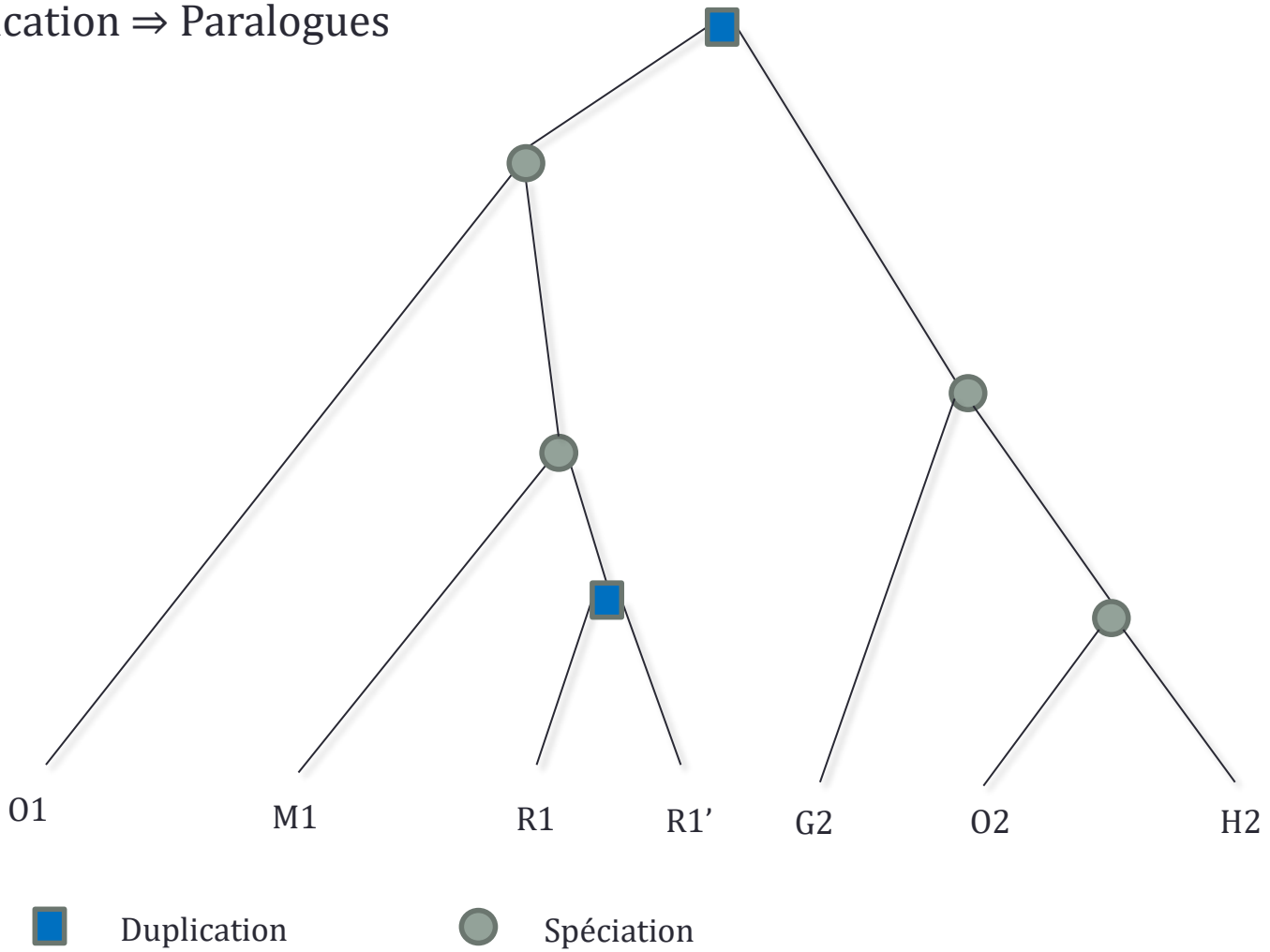
Paralogues si leur **ancêtre commun le plus récent** a subit une **duplication**.

Lca = spéciation ⇒ Orthologues

Lca = duplication ⇒ Paralogues

(*lca = least common ancestor*)

O1 et M1?

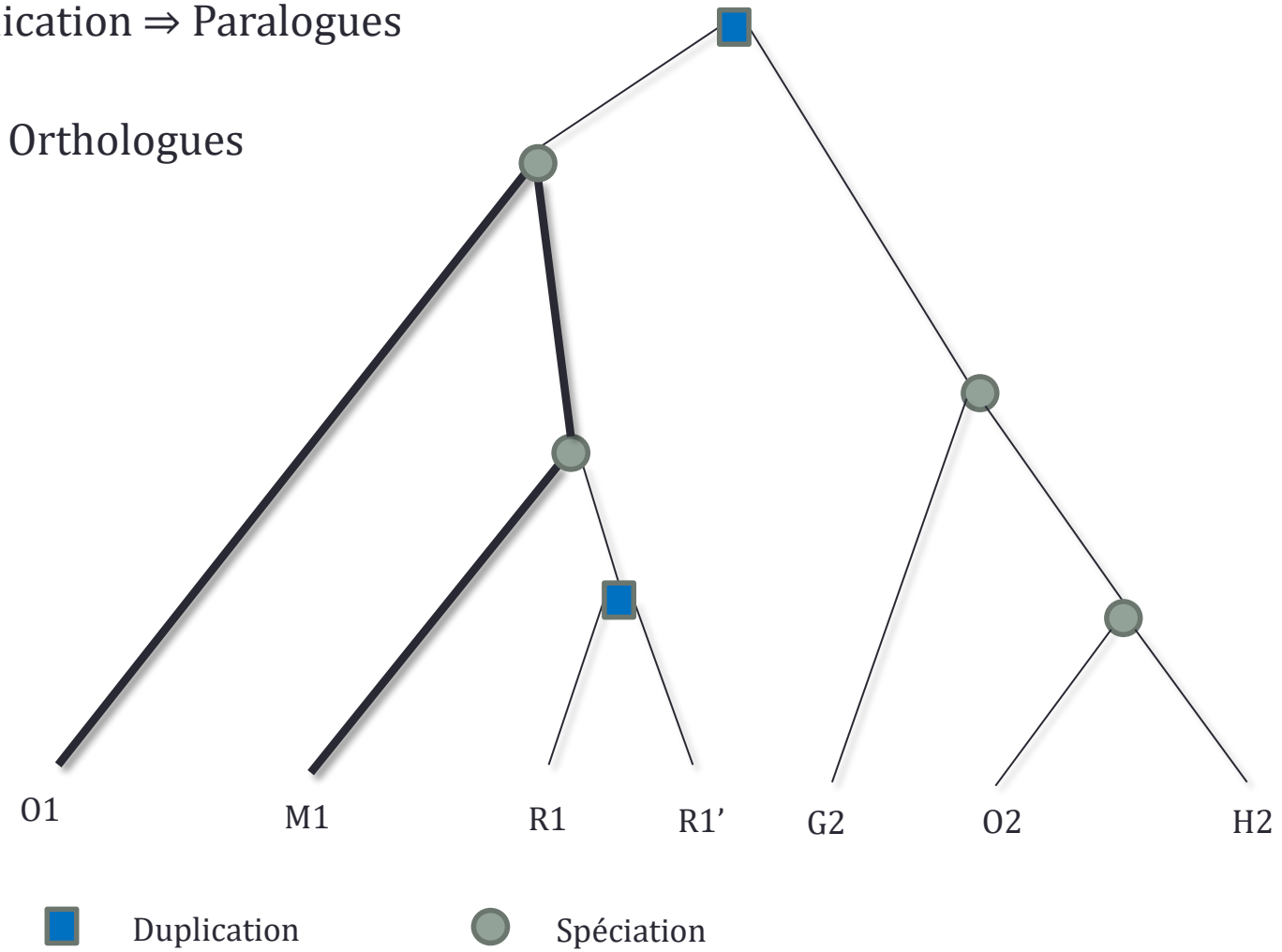


Lca = spéciation ⇒ Orthologues

Lca = duplication ⇒ Paralogues

O1 et M1? Orthologues

(*lca = least common ancestor*)

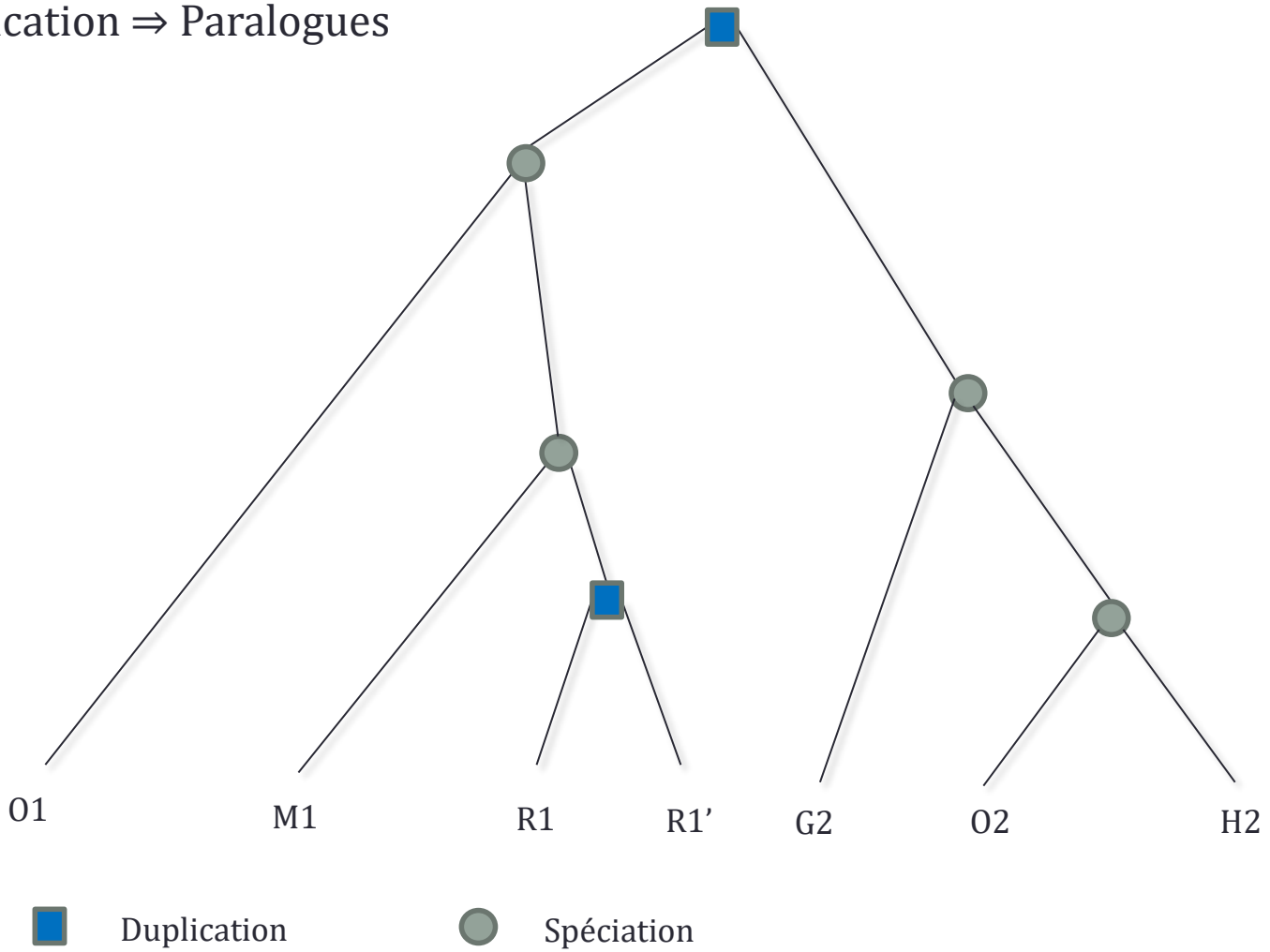


Lca = spéciation ⇒ Orthologues

Lca = duplication ⇒ Paralogues

(*lca = least common ancestor*)

M1 et G2?

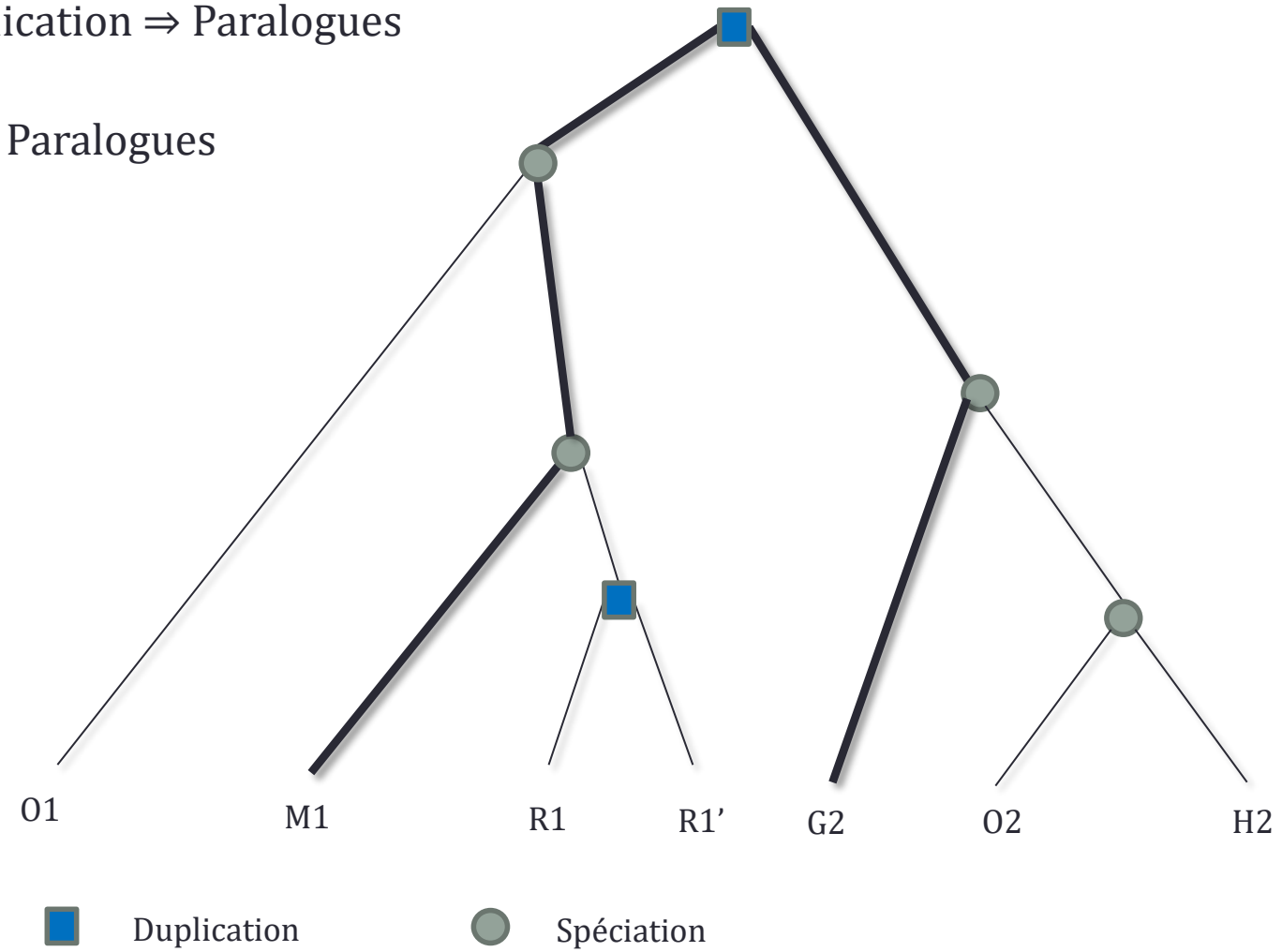


Lca = spéciation ⇒ Orthologues

Lca = duplication ⇒ Paralogues

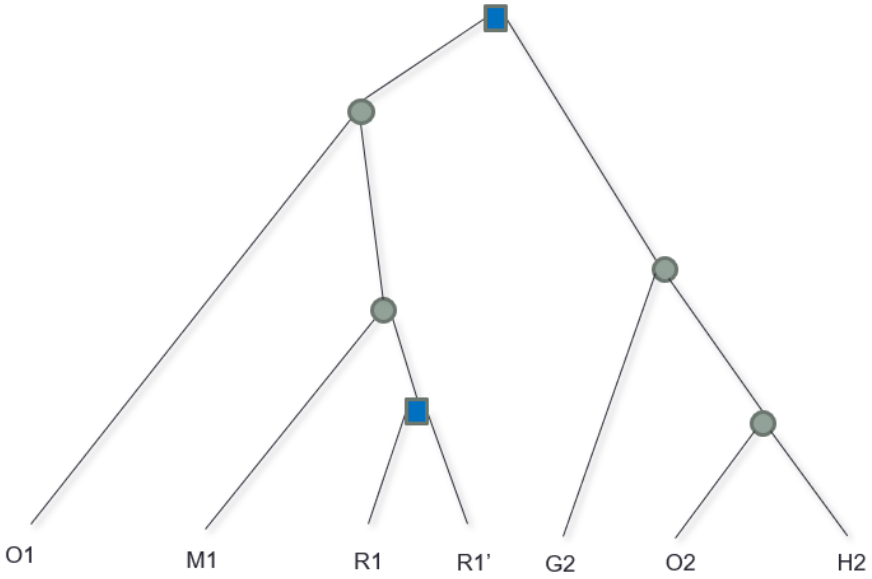
(*lca = least common ancestor*)

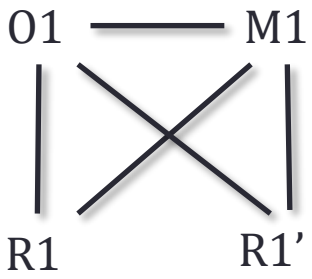
M1 et G2? Paralogues

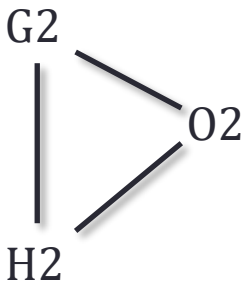


Graphe d'orthologie

- Sommets = gènes (feuilles de l'arbre)
- Arête = orthologues (non-arête = paralogue)





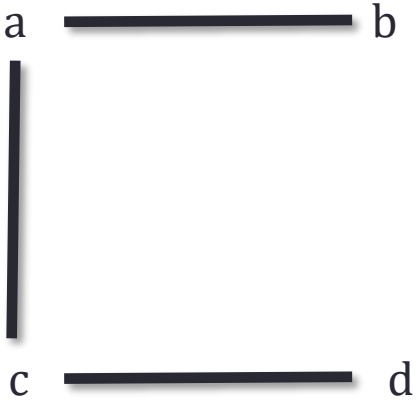


Graphe d'orthologie

- Quelques questions sur les graphes d'orthologie:
  - Est-ce que **tous les graphes** sont des graphes d'orthologie?
  - Est-ce qu'un graphe d'orthologie correspond à un **arbre unique**?
  - Si on n'a que le graphe, **comment reconstruire** un tel arbre?

Graphe d'orthologie

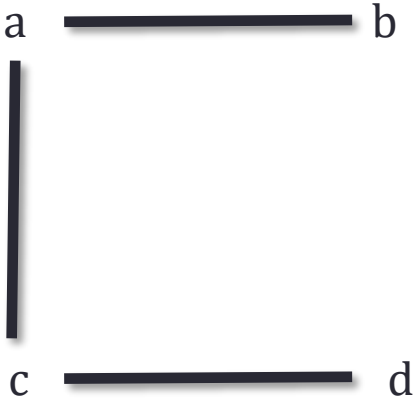
- La plupart des logiciels produisent en sortie un graphe d'orthologie.



= Orthologues (liés par spéciation)

Graphe d'orthologie

- La plupart des logiciels produisent en sortie un graphe d'orthologie.

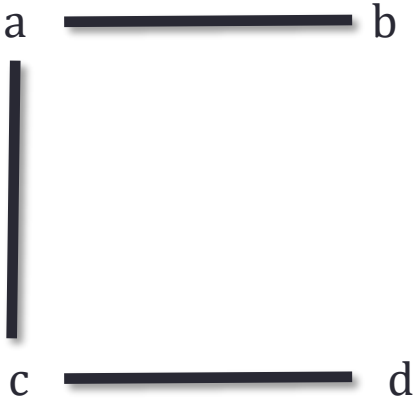


Bon? Mauvais?
Comment savoir???

= Orthologues (liés par spéciation)

Graphe d'orthologie

- Si la machine dit vrai, on devrait pouvoir construire un arbre de gènes **qui exhibe les relations de lca** données.



Bon? Mauvais?
Comment savoir???

= Orthologues (liés par spéciation)

Et à quoi ça sert?

- À **prédire les fonctionnalités** des gènes.
- "Conjecture des orthologues" *[Koonin, 2005]*
  - Les gènes **orthologues** (liés par spéciation) ont tendance à être **similaires** du point de vue des fonctions et des séquences d'ADN.
  - Les gènes **paralogues** (liés par duplication) ont tendance à **diverger**.

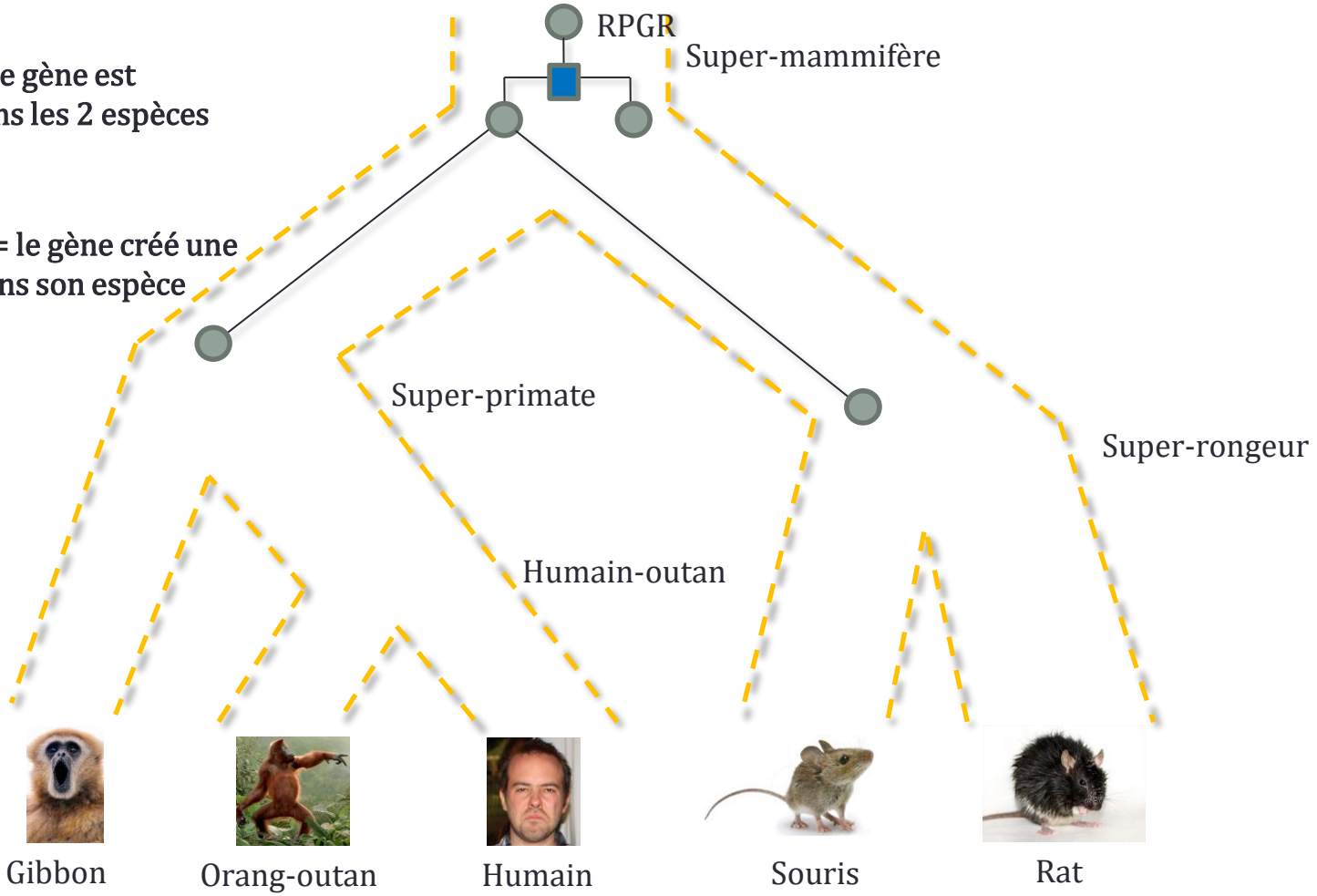
Et à quoi ça sert?

- À **prédire les fonctionnalités** des gènes.
- "Conjecture des orthologues" *[Koonin, 2005]*
  - Les gènes **orthologues** (liés par spéciation) ont tendance à être **similaires** du point de vue des fonctions et des séquences d'ADN.
  - Les gènes **paralogues** (liés par duplication) ont tendance à **diverger**.

- Connaître la fonction d'un gène = connaître la fonction de tous ses gènes orthologues.
  - Si on peut trouver ces orthologues, évite d'avoir à tout tester par expérimentation biologique.

Spéciation = le gène est « envoyé » dans les 2 espèces descendants

Duplication = le gène créé une 2ème copie dans son espèce



Et qui ça intéresse?

- Il existe environ une **cinquantaine de logiciels** de prédiction d'orthologues toujours maintenus.
  - Les plus populaires: COG, OrthoMCL, proteinortho, OMA, ...
  - Fournissent aussi une base de données (BD) d'orthologues.
  - La plupart basés sur des **heuristiques**.

- Consortium ***Quest for Orthologs***.
  - Objectif: évaluer et améliorer les méthodes émergentes de prédiction d'orthologie afin d'obtenir des données d'orthologies précises.
    - e.g. orthoMCL, OMA

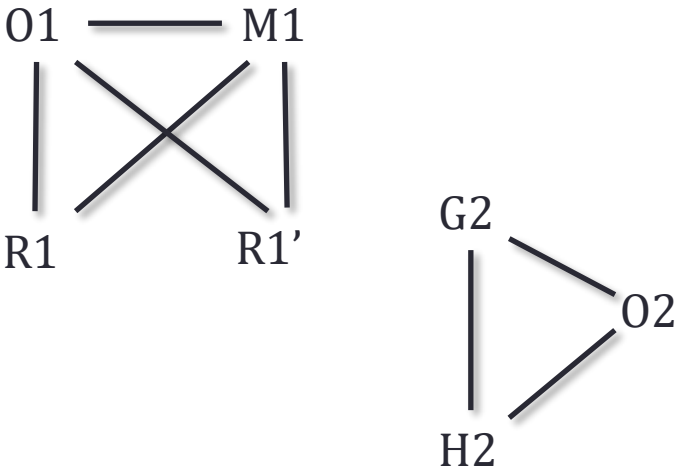
Caractérisation des graphes d'orthologie

Graphe d'orthologie

- **Définition :** G est une graphe d'orthologie s'il existe un arbre T dont les feuilles sont $V(G)$, et tel que $uv \in E(G)$ ssi $lca_T(u, v)$ est marqué "spéciation".

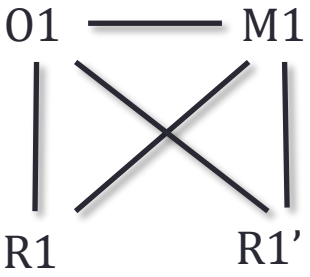
Graphe d'orthologie

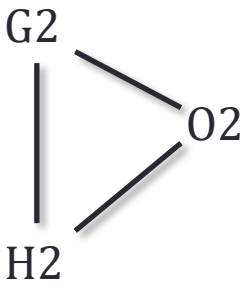
- **Définition :** G est une graphe d'orthologie s'il existe un arbre T dont les feuilles sont $V(G)$, et tel que $uv \in E(G)$ ssi $lca_T(u, v)$ est marqué "spéciation".

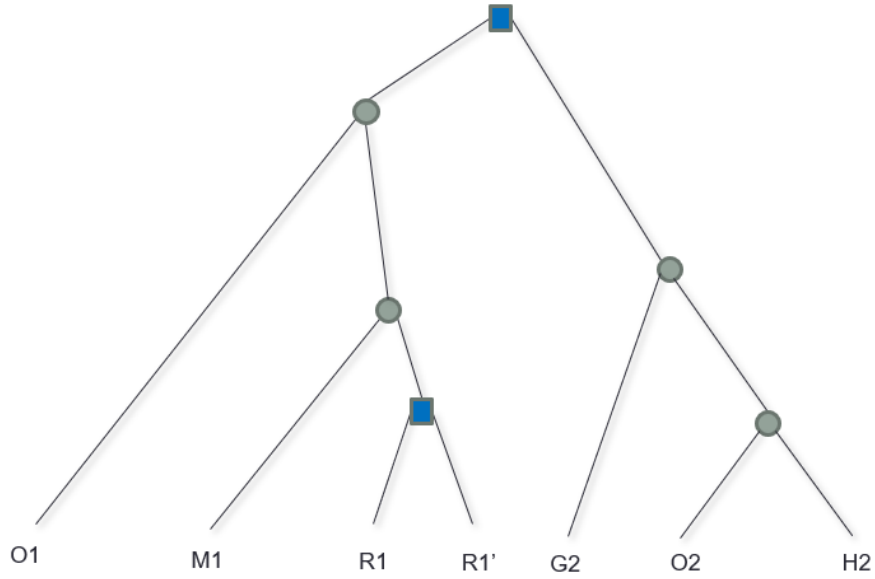


Graphe d'orthologie

- **Définition :** G est une graphe d'orthologie s'il existe un arbre T dont les feuilles sont $V(G)$, et tel que $uv \in E(G)$ ssi $lca_T(u, v)$ est marqué "spéciation".







Graphe d'orthologie

- **Définition :** G est une graphe d'orthologie s'il existe un arbre T dont les feuilles sont $V(G)$, et tel que $uv \in E(G)$ ssi $lca_T(u, v)$ est marqué "spéciation".

- Est-ce que tous les graphes sont des graphes d'orthologie?

Graphe d'orthologie

- **Définition :** G est une graphe d'orthologie s'il existe un arbre T dont les feuilles sont $V(G)$, et tel que $uv \in E(G)$ ssi $lca_T(u, v)$ est marqué "spéciation".

- Est-ce que tous les graphes sont des graphes d'orthologie?
- NON

Graphe d'orthologie

- **Théorème** : G est un graphe d'orthologie si et seulement si il n'a pas de P_4 induit.

Graphe d'orthologie

- **Théorème** : G est un graphe d'orthologie si et seulement si il n'a pas de P_4 induit.

- (P_4 induit = chemin avec 4 sommets, sans raccourcis)

Graphe d'orthologie

- **Théorème** : G est un graphe d'orthologie si et seulement si il n'a pas de P_4 induit.

- À prouver

⇒ Si G est un graphe d'orthologie, alors il n'a pas de P_4 induit.

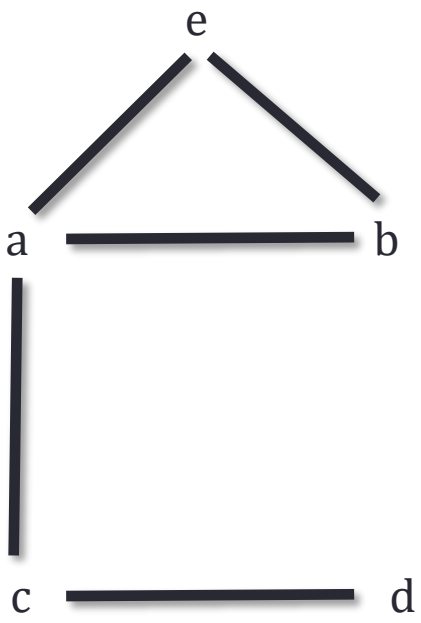
⇐ Si G n'a pas de P_4 induit, c'est un graphe d'orthologie.

Graphe d'orthologie

⇒ Si G est un graphe d'orthologie, alors il n'a pas de P_4 induit.

i.e. si G a un P_4 induit, alors G n'est pas un graphe d'orthologie.

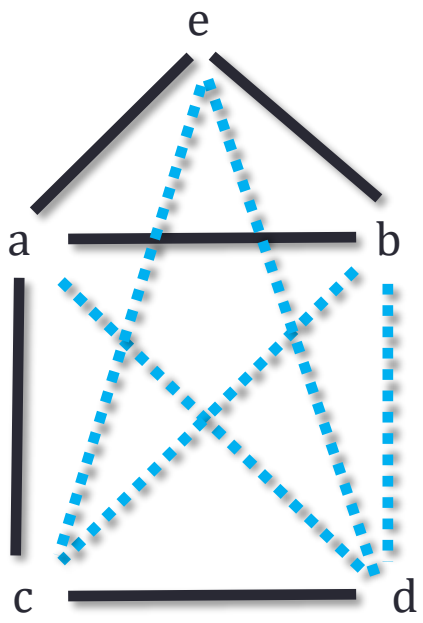
Graphe d'orthologie



= Orthologues (liés par spéciation)

= Paralogues (liés par duplication)

Graphe d'orthologie

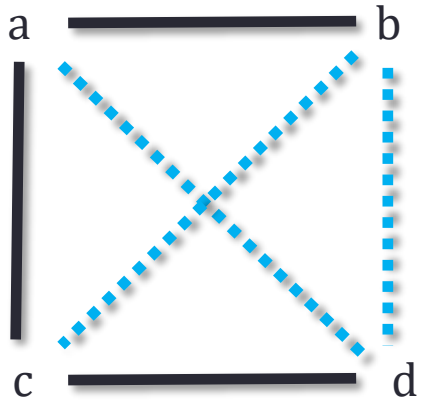


= Orthologues (liés par spéciation)

= Paralogues (liés par duplication)

Graphe d'orthologie

Un P4 induit!!

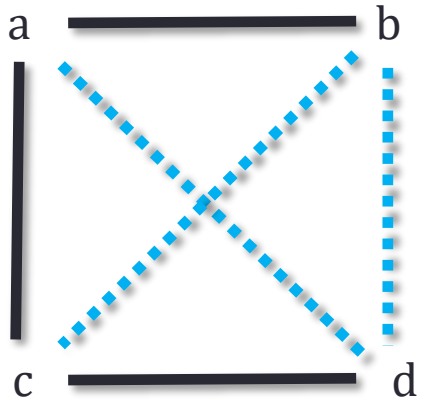


= Orthologues (liés par spéciation)

= Paralogues (liés par duplication)

Graphe d'orthologie

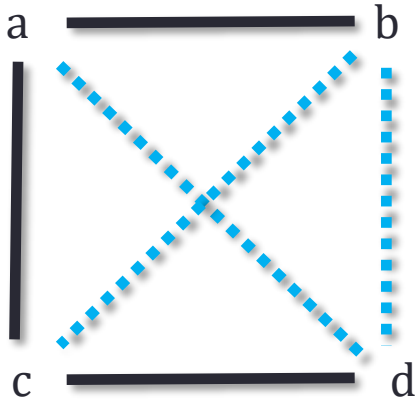
Un P4 induit!!



Si G était un graphe d'orthologie, on pourrait trouver un arbre dans lequel
lca(a, b) = Spec
lca(a, c) = Spec
lca(c, d) = Spec
lca(a, d) = Dup
lca(b, c) = Dup
lca(b, d) = Dup

= Orthologues (liés par spéciation)

= Paralogues (liés par duplication)

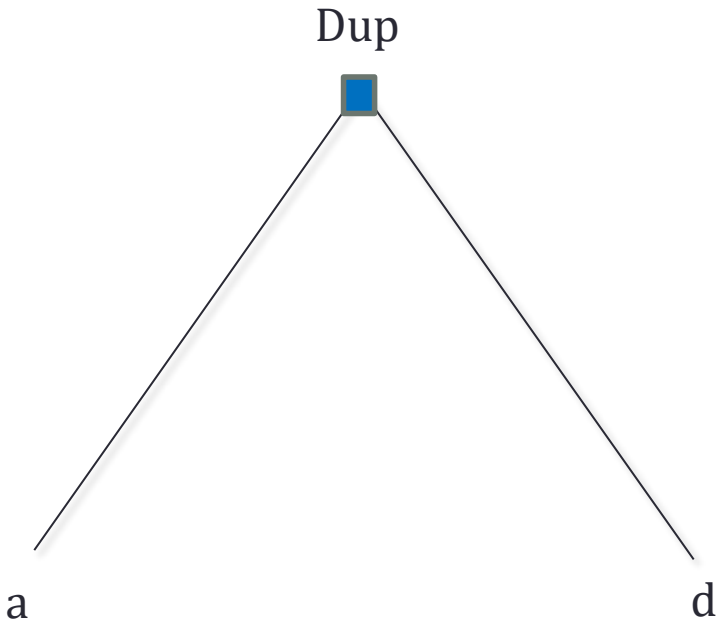


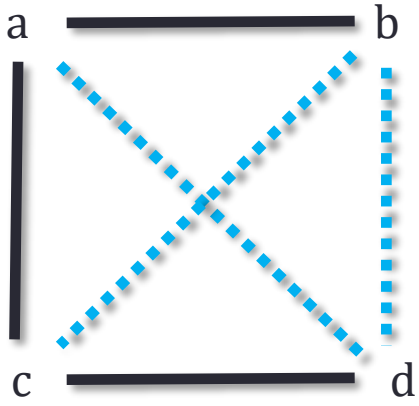
= Orthologues (spéciation)

= Paralogues (duplication)

Si G était un graphe d'orthologie, on pourrait trouver un arbre dans lequel

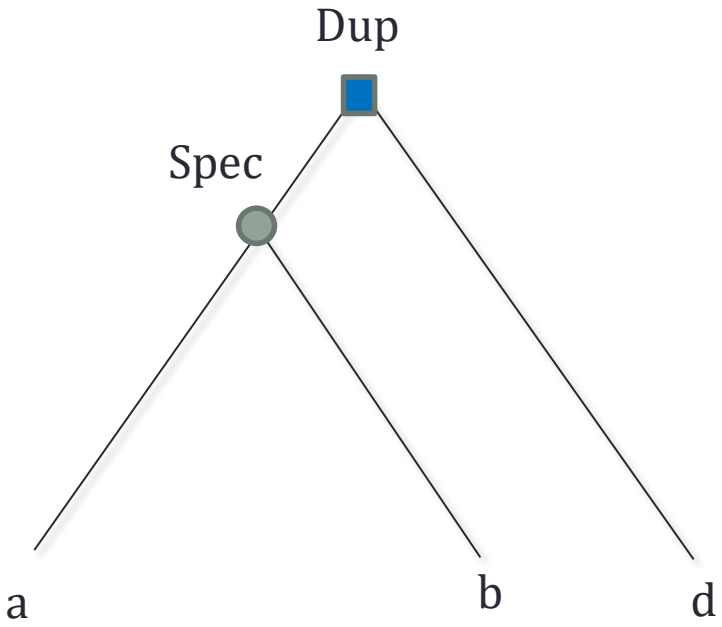
lca(a, b) = Spec
lca(a, c) = Spec
lca(c, d) = Spec
lca(a, d) = Dup
lca(b, c) = Dup
lca(b, d) = Dup



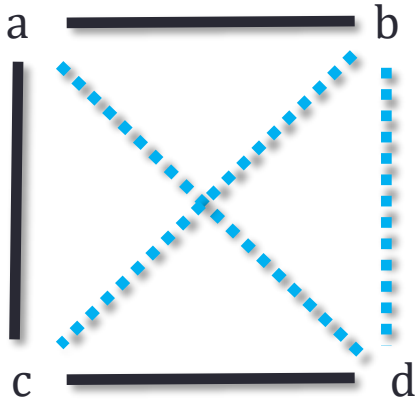


= Orthologues (spéciation)

= Paralogues (duplication)

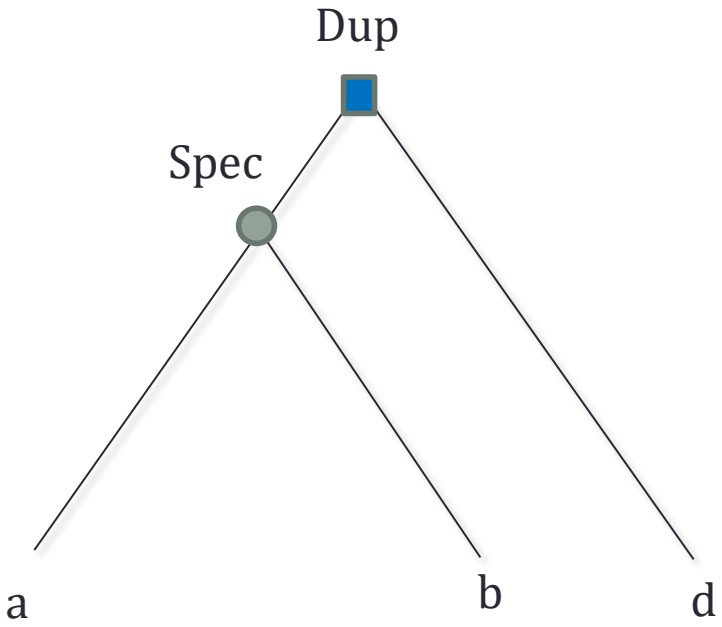


Si G était un graphe d'orthologie, on pourrait trouver un arbre dans lequel
lca(a, b) = Spec
lca(a, c) = Spec
lca(c, d) = Spec
lca(a, d) = Dup
lca(b, c) = Dup
lca(b, d) = Dup

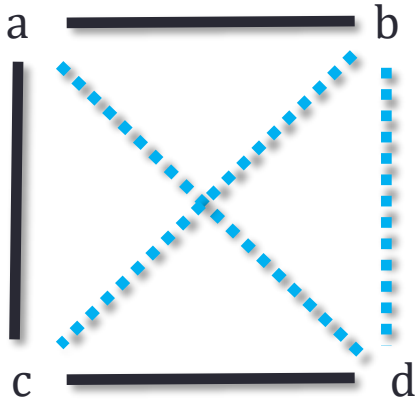


= Orthologues (spéciation)

= Paralogues (duplication)

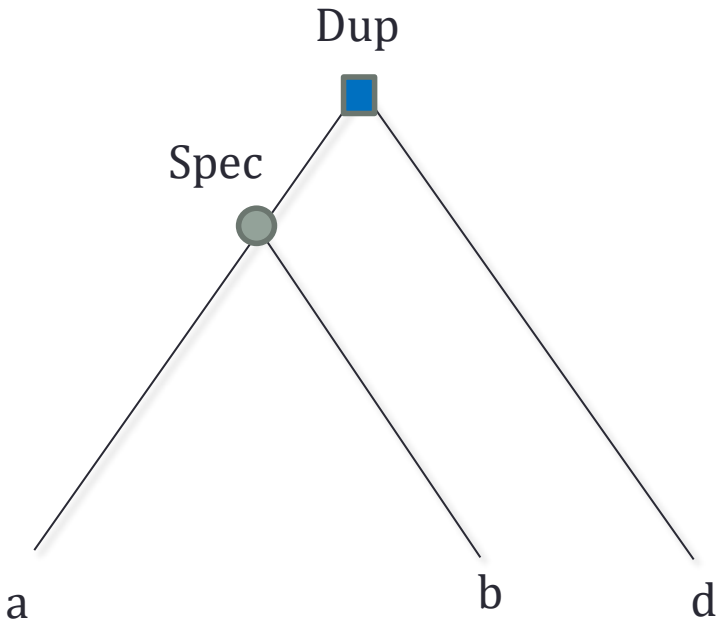


Si G était un graphe d'orthologie, on pourrait trouver un arbre dans lequel
lca(a, b) = Spec
lca(a, c) = Spec
lca(c, d) = Spec
lca(a, d) = Dup
lca(b, c) = Dup
lca(b, d) = Dup

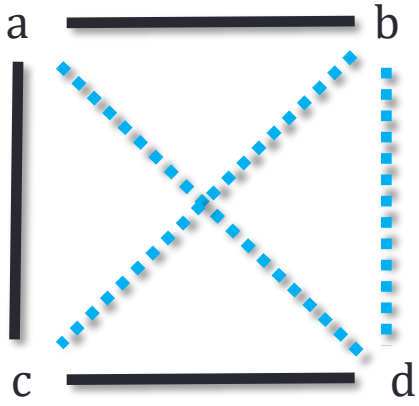


= Orthologues (spéciation)

= Paralogues (duplication)

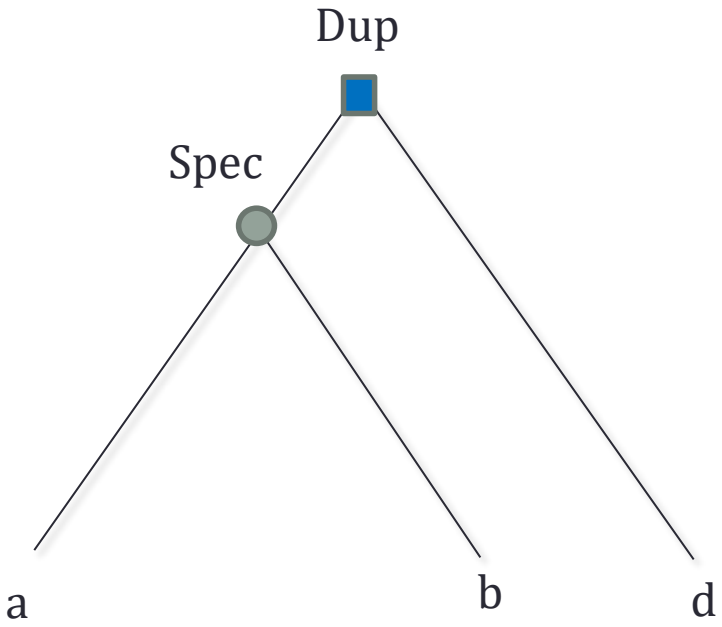


Si G était un graphe d'orthologie, on pourrait trouver un arbre dans lequel
lca(a, b) = Spec
lca(a, c) = Spec
lca(c, d) = Spec
lca(a, d) = Dup
lca(b, c) = Dup
lca(b, d) = Dup

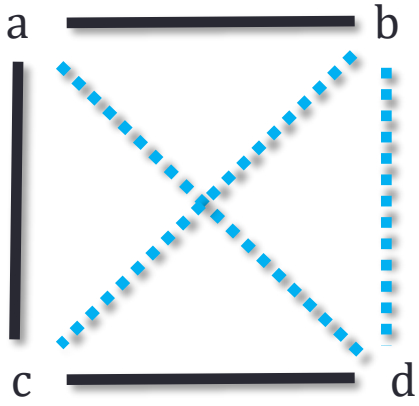


= Orthologues (spéciation)

= Paralogues (duplication)

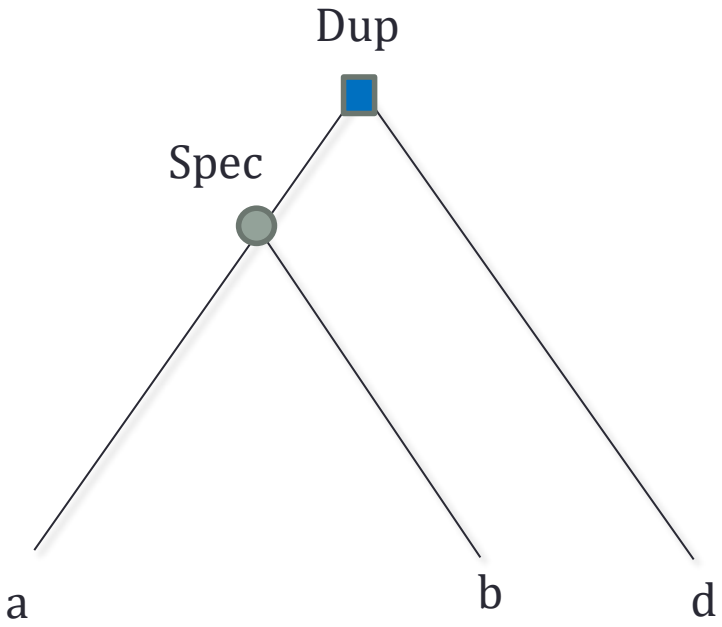


Si G était un graphe d'orthologie, on pourrait trouver un arbre dans lequel
lca(a, b) = Spec
lca(a, c) = Spec
lca(c, d) = Spec
lca(a, d) = Dup
lca(b, c) = Dup
lca(b, d) = Dup



= Orthologues (spéciation)

= Paralogues (duplication)



Si G était un graphe d'orthologie, on pourrait trouver un arbre dans lequel
lca(a, b) = Spec
lca(a, c) = Spec
lca(c, d) = Spec
lca(a, d) = Dup
lca(b, c) = Dup
lca(b, d) = Dup

Graphe d'orthologie

⇒ Si G est un graphe d'orthologie, alors il n'a pas de P_4 induit.

"Preuve" : par force brute, on peut voir qu'il n'existe pas d'arbre capable d'exhiber les relations de lca données par un P_4 induit.

Graphe d'orthologie

- **Théorème** : G est un graphe d'orthologie si et seulement si il n'a pas de P_4 induit.

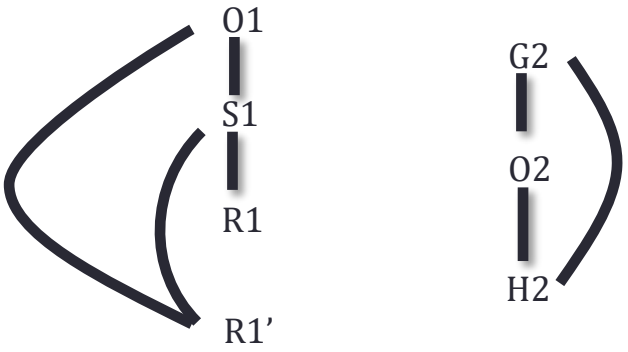
- À prouver

~~⇒ Si G est un graphe d'orthologie, alors il n'a pas de P_4 induit.~~

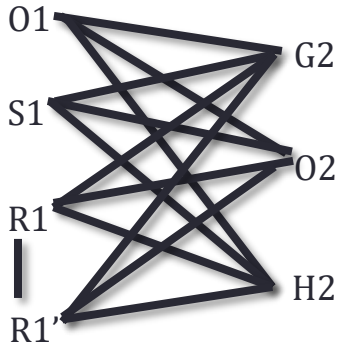
⇐ Si G n'a pas de P_4 induit, c'est un graphe d'orthologie.

- **Définition**: soit G un graphe. Le **complément** de G, dénoté $\bar{G}$, est le graphe dans lequel les arêtes sont les non-arêtes de G.

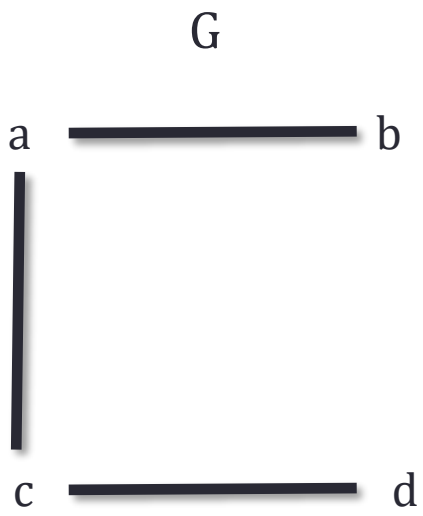
G

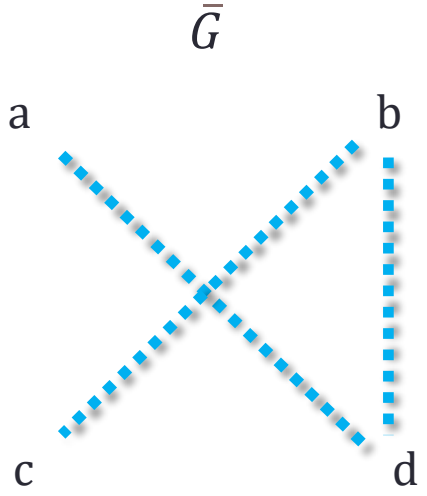


$\bar{G}$



- **Définition**: soit G un graphe. Le **complément** de G, dénoté $\bar{G}$, est le graphe dans lequel les arêtes sont les non-arêtes de G.





- **Théorème [Corneil, 1985]**: soit G un graphe sans P_4 induit avec au moins 2 sommets. Alors G est déconnecté, ou son complément $\bar{G}$ est déconnecté.

- **Théorème [Corneil, 1985]**: soit G un graphe sans P_4 induit avec au moins 2 sommets. Alors G est déconnecté, ou son complément $\bar{G}$ est déconnecté.

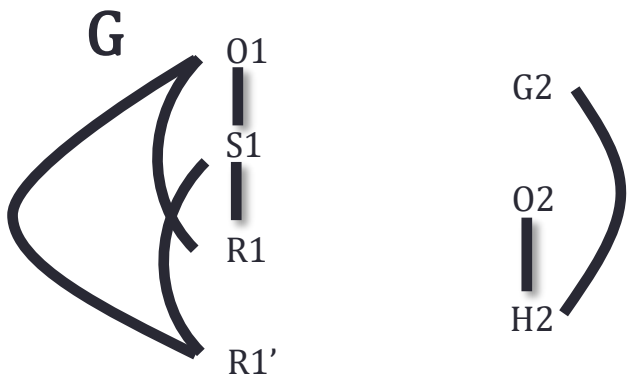
- **Théorème [Corneil, 1985]**: soit G un graphe sans P_4 induit avec au moins 2 sommets. Alors G est déconnecté, ou son complément $\bar{G}$ est déconnecté.

- **Théorème [Corneil, 1985]**: soit G un graphe sans P_4 induit avec au moins 2 sommets. Alors G est déconnecté, ou son complément $\bar{G}$ est déconnecté.

- **Théorème [Corneil, 1985]**: soit G un graphe sans P_4 induit avec au moins 2 sommets. Alors G est déconnecté, ou son complément $\bar{G}$ est déconnecté.

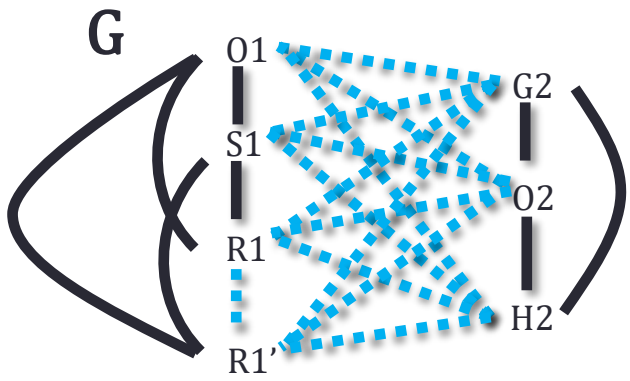
- Revenons à l’orthologie.
- À prouver:

⇐ Si G n'a pas de P_4 induit, c'est un graphe d'orthologie.

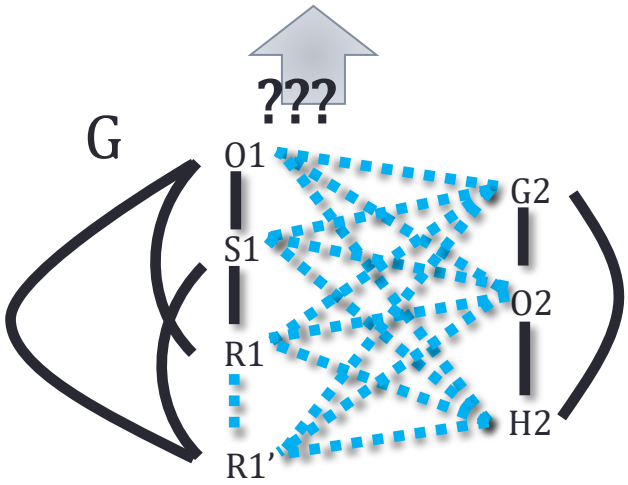


- Revenons à l’orthologie.
- À prouver:

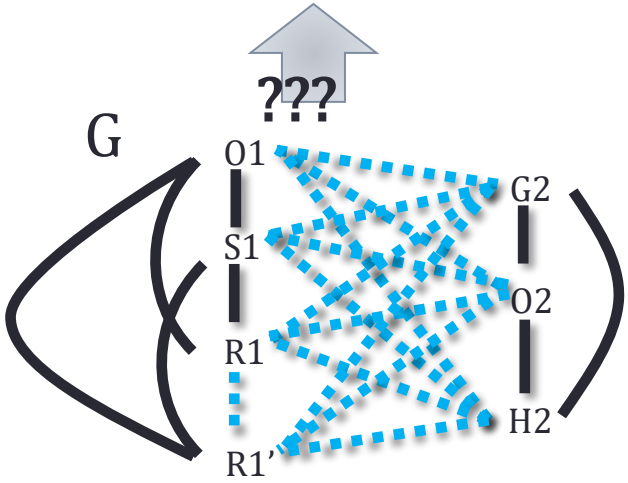
⇐ Si G n'a pas de P_4 induit, c'est un graphe d'orthologie.



Arbre pour G avec les bonnes relations de lca?

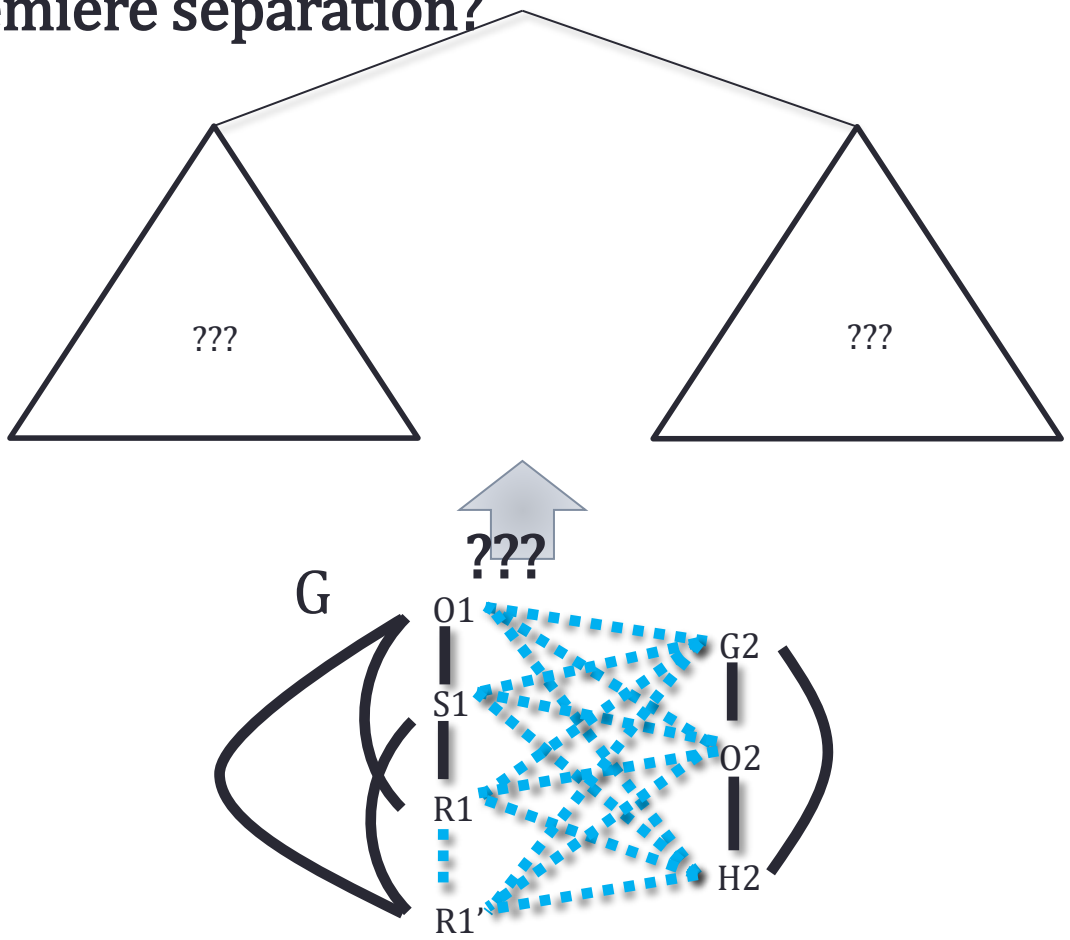


Arbre pour G avec les bonnes relations de lca?

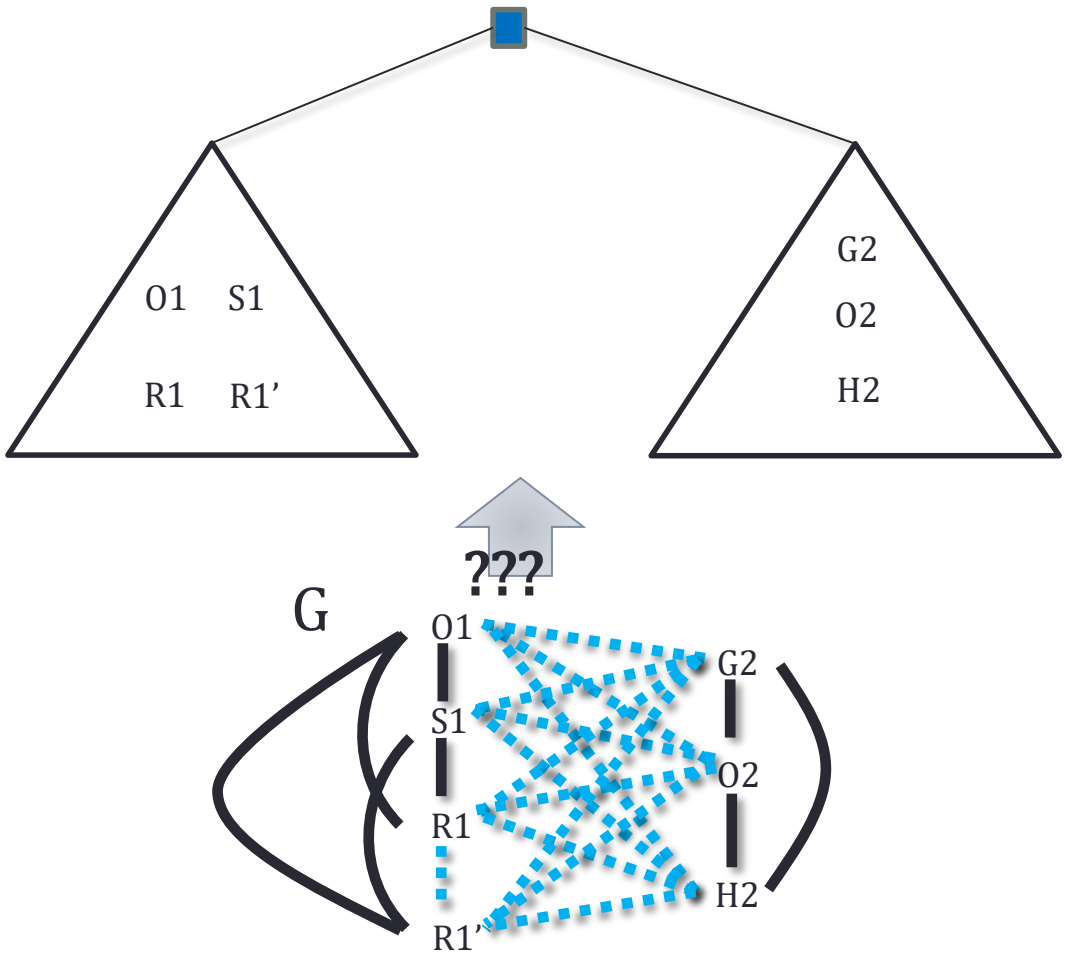


- **Thm [Corneil, 1985]**: si G n'a pas de P_4 induit, alors G ou son complément est déconnecté.

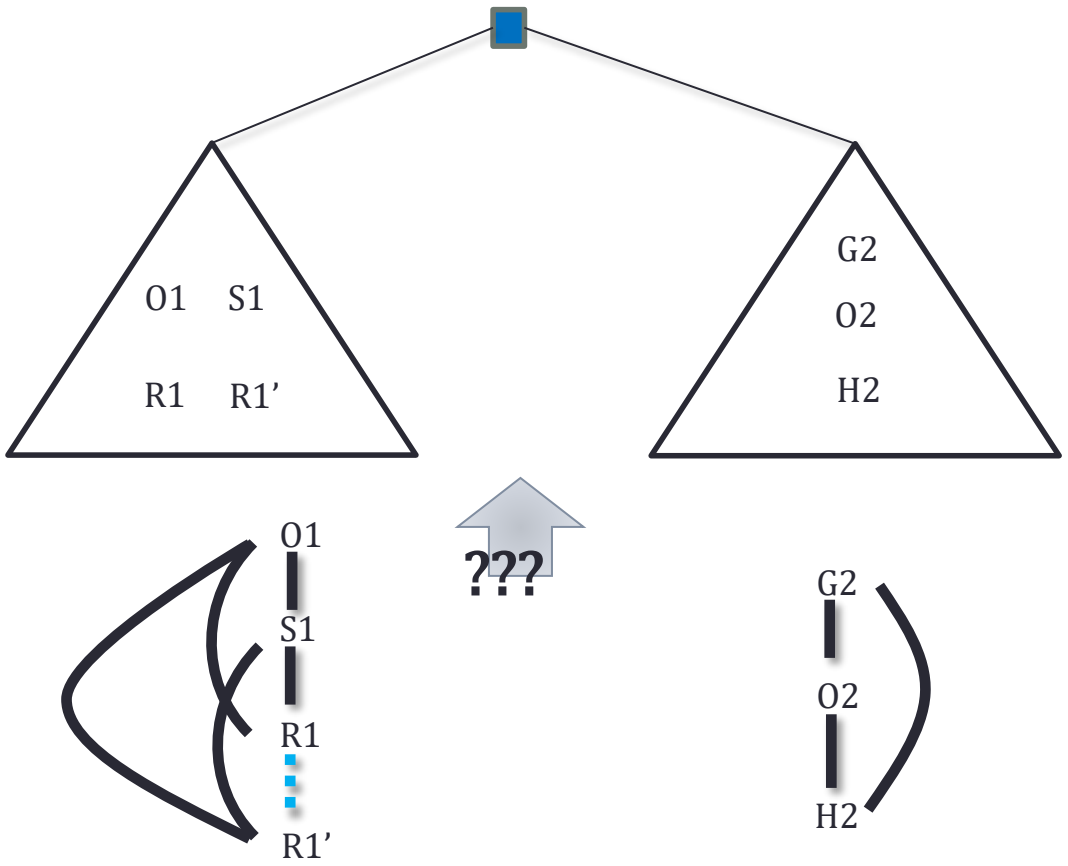
Commençons par la racine? Quelle est la première séparation?



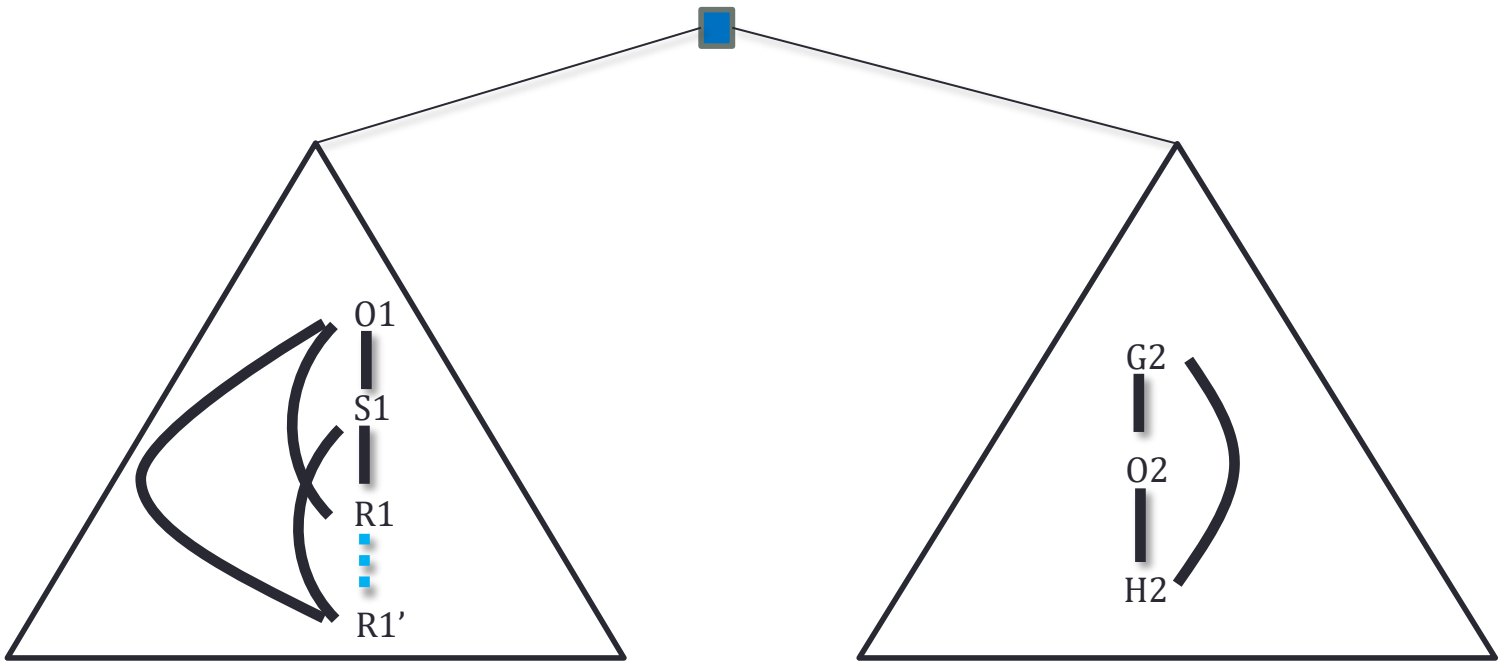
- **Thm [Corneil, 1985]**: si G n'a pas de P_4 induit, alors G ou son complément est déconnecté.



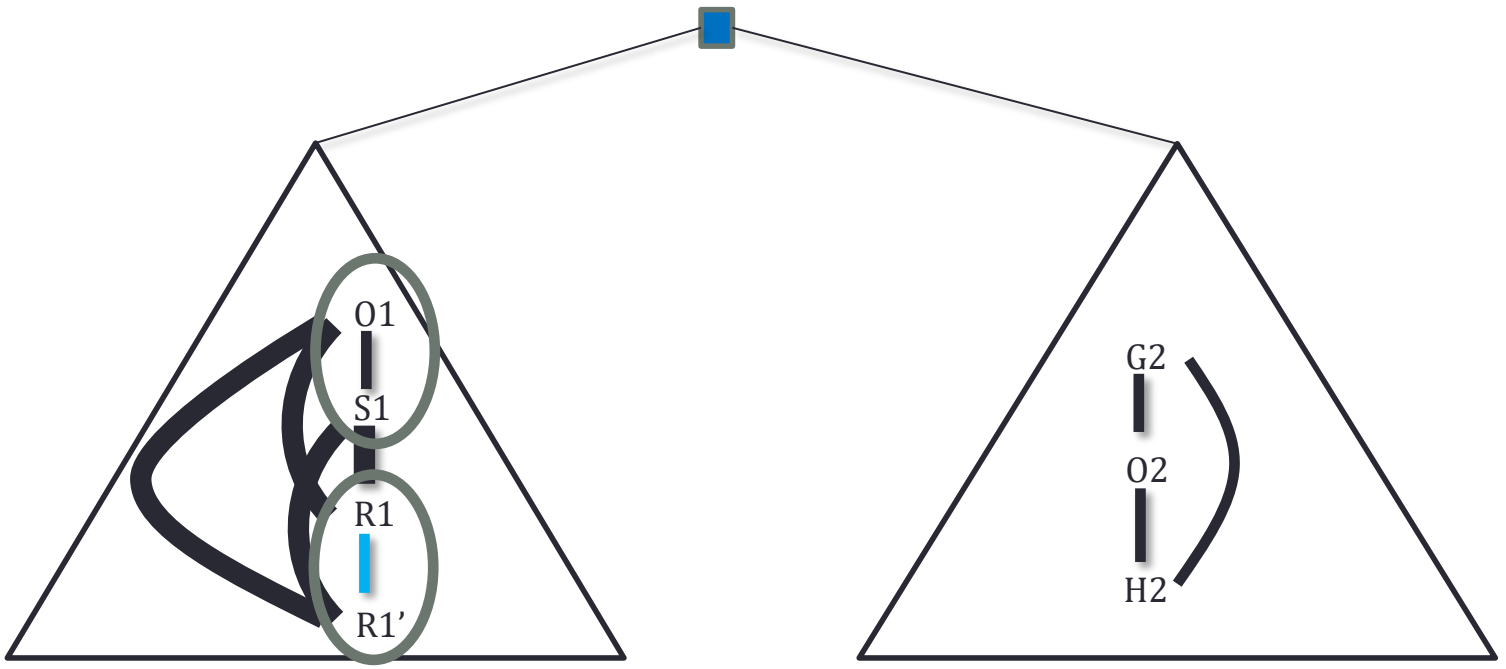
- **Thm [Corneil, 1985]**: si G n'a pas de P_4 induit, alors G ou son complément est déconnecté.



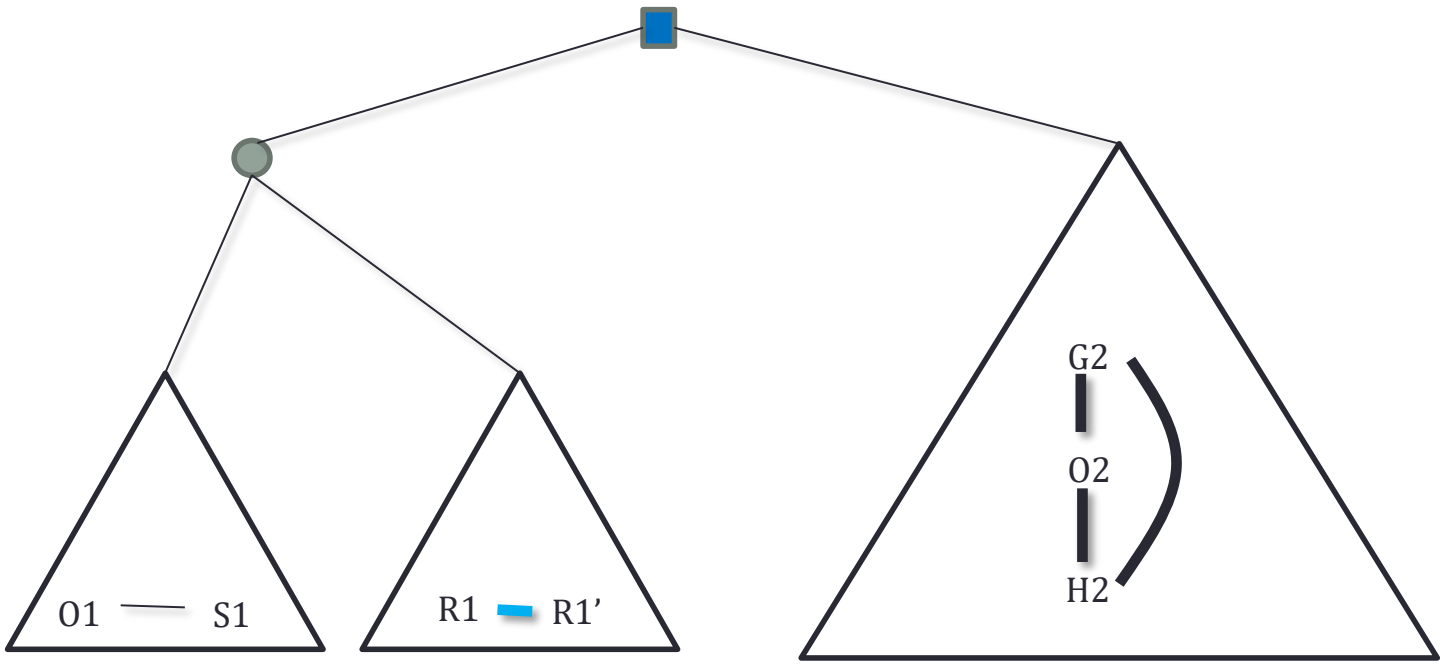
- **Thm [Corneil, 1985]**: si G n'a pas de P_4 induit, alors G ou son complément est déconnecté.



- **Thm [Corneil, 1985]**: si G n'a pas de P_4 induit, alors G ou son complément est déconnecté.



- **Thm [Corneil, 1985]**: si G n'a pas de P_4 induit, alors G ou son complément est déconnecté.



et ainsi de suite…

- **Thm [Corneil, 1985]**: si G n'a pas de P_4 induit, alors G ou son complément est déconnecté.

Validité de relations

Du point de vue algorithmique, vérifier si un graphe contient un P_4:

- peut se faire en temps $O(n^4)$ naïvement (n = nombre de noeuds).
- peut se faire en temps $O(n)$ de façon intelligente (mais compliquée).
- Dans *[Lafond & El-Mabrouk, 2014]*, implémentation en temps $O(n^2)$.

Validité de relations

Du point de vue algorithmique, vérifier si un graphe contient un P_4:

- peut se faire en temps $O(n^4)$ naïvement (n = nombre de noeuds).
- peut se faire en temps $O(n)$ de façon intelligente (mais compliquée).
- Dans *[Lafond & El-Mabrouk, 2014]*, implémentation en temps $O(n^2)$.

- Expérimentation sur 265 familles de gènes avec le logiciel proteinortho:
  - Avec paramètres par défaut, **90% des graphes en sortie sont invalides.**
  - Avec paramètres optimaux, **70% des graphes en sortie sont invalides.**

Un peu de littérature

Théorème des P4 [Hernandez & Rosalez et al., JoMB 2013]

Théorème des P4 [Lafond & El-Mabrouk, BMC Bioinf 2014]

Résultats algorithmiques

- Si certaines relations sont inconnues (orthologue ou paralogue), peut-on les choisir sans créer de P4
- [Lafond & El-Mabrouk, BMC Bioinf 2014]
- Si le graphe a des P4, modifier un # minimum d'arêtes pour qu'il n'en ait plus. NP-complet, pas de c-approx.
- [Dondi, Lafond & El-Mabrouk, AMB 2017]

Un peu de littérature

Théorème des P4 [Hernandez & Rosalez et al., JoMB 2013]

Théorème des P4 [Lafond & El-Mabrouk, BMC Bioinf 2014]

Un peu de littérature

Résultats algorithmiques

- Si certaines **relations sont inconnues** (orthologue ou paralogue), peut-on les choisir sans créer de P_4
  - Faisable en temps polynomial [Lafond & El-Mabrouk, BMCBio 2014]

- Si le graphe a des P_4, **modifier un # minimum d'arêtes** pour qu'il n'en ait plus.
  - NP-complet, pas de c-approx. [Dondi, Lafond & El-Mabrouk, AMB 2017]

- Correction de graphes de **degré borné**
  - Algos FPT [Lopez & Lafond., JCBC 2021]

Autres modèles de graphes

Autres types de **relation de lca**

- M Geiß et al., Best match graphs, JoMB 2016
- Hellmuth & al., Inferring phylogenetic trees from the knowledge of rare evolutionary events, JoMB 2018

Avec événements de **transfert de gènes**

- M Geiß & al., Reconstructing gene trees from Fitch's xenology relation, JoMB, 2018
- M Lafond, M Hellmuth, Reconstruction of time-consistent species trees, AMB 2021
- D Schaller & al, Indirect identification of horizontal gene transfer, JoMB 2021

Arêtes qui experiment des **contraintes de distances**

- PF Stadler & al., From pairs of most similar sequences to phylogenetic best matches, AMB 2020
- M Lafond, Recognizing k-leaf powers in polynomial time, for constant k, SODA 2022

…

Conclusion

- Certaines relations évolutives peuvent être représentées sous la forme d'un graphe.
- Donne lieu à des problèmes de caractérisation de graphe + problèmes algorithmiques sur ces graphes.
- Meilleure compréhension de l'évolution et des conséquences de nos modèles.