# Méthodes de construction et de comparaison d'arbres retraçant l'évolution des espèces depuis des données phylogénomiques

**Yimin SHEN**

Soutenu par le projet Phylariane – http://www.lirmm.fr/phylariane/
ANR-08-EMER-011-01

Tuteurs du stage en laboratoire :
- **Laurent BREHELIN**
- **Vincent BERRY**
- **Emmanuel DOUZERY**

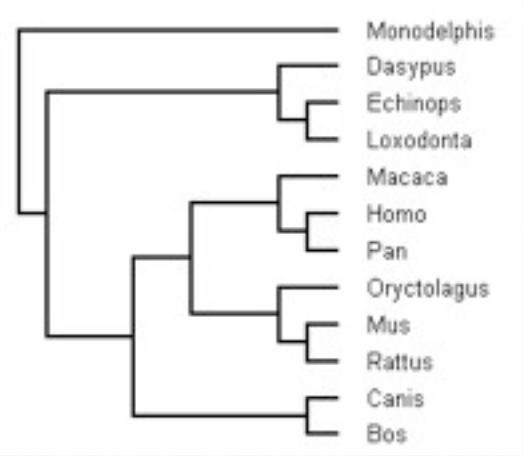
Tuteur du stage à l'université :
- **Vincent RANWEZ**

Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM)

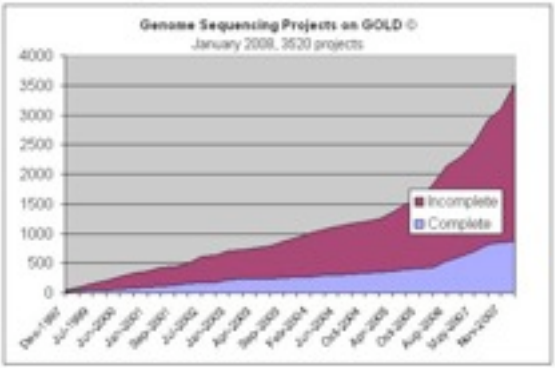
Master Mention Informatique
Spécialité Intégration de Compétences
Parcours Bio-informatique
Université Montpellier 2

# Notions

- La **phylogénie** est la représentation sous forme arborée de la formation et de l'évolution des organismes vivants au cours du temps. Une phylogénie indique entre autres les parentés entre groupes d'espèces




- La **phylogénomique** désigne la rencontre de deux champs de recherche : la phylogénie et la génomique. C'est d'analyse impliquant des données génomiques et de l'inférence évolutive, en particulier la reconstruction phylogénétique.




  - Il y a de plus en plus de génomes qui sont séquencés. Donc, on a de plus en plus d'arbres de gènes, et une question importante est de savoir comment obtenir l'arbre des espèces depuis cette masse de données.

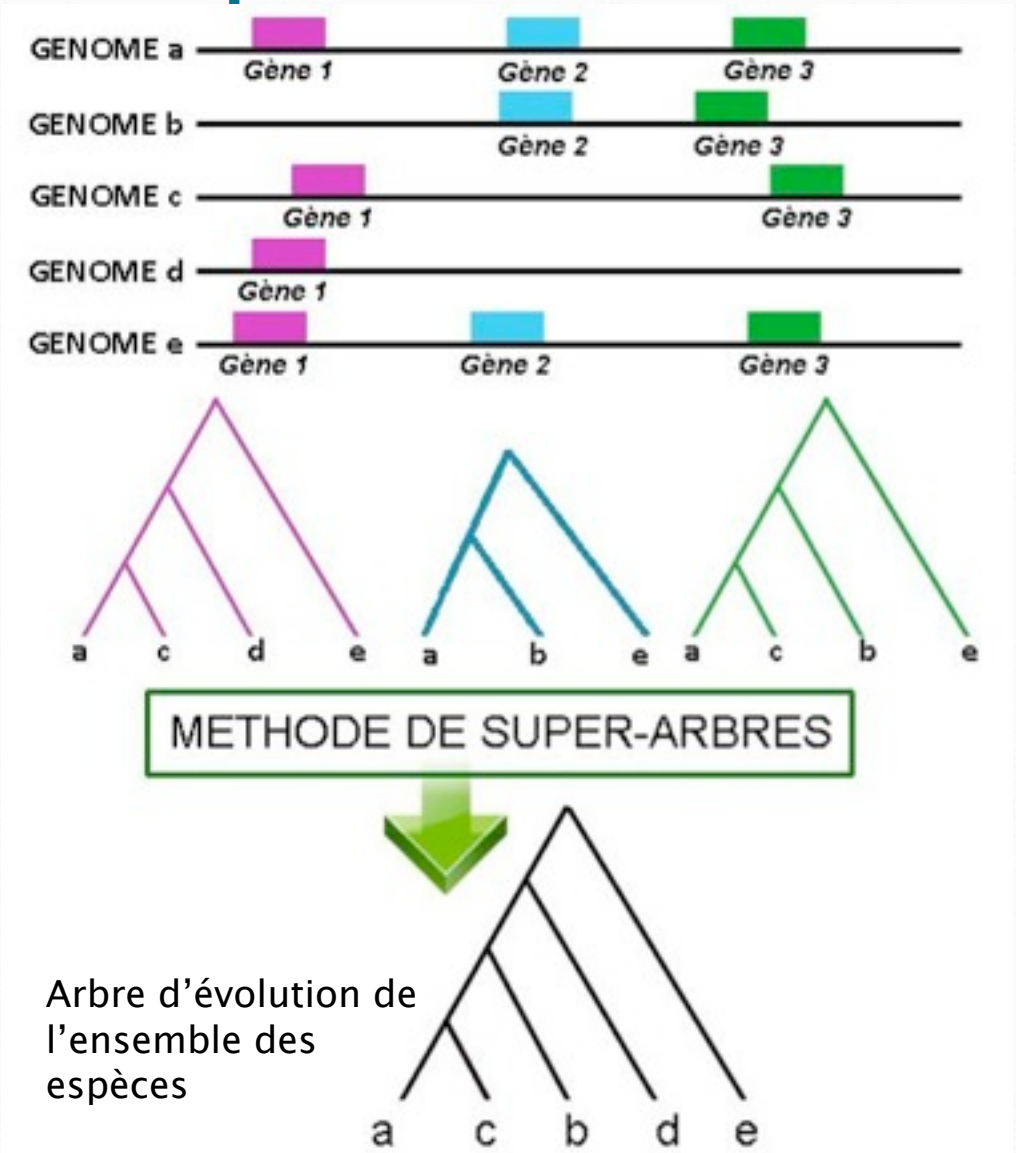
# Méthodes de super-arbres

# Méthodes de super-arbres

- Que-ce que c'est?

# Méthodes de super-arbres

- Que-ce que c'est?




Arbre d'évolution de l'ensemble des espèces

# Méthodes de super-arbres

- Que-ce que c'est?

- Méthodes de consensus
  - Strict Consensus Supertree [Gordon 86]
  - Reduced Consensus Supertree [Wilkinson Thorley 02]
  - Maximum Agreement Supertree et Maximum Compatible Supertree [Berry et al. 04]

# Méthodes de super-arbres

- Que-ce que c'est?

- Méthodes de consensus
  - Strict Consensus Supertree [Gordon 86]
  - Reduced Consensus Supertree [Wilkinson Thorley 02]
  - Maximum Agreement Supertree et Maximum Compatible Supertree [Berry et al. 04]

- Méthodes d'optimisation
  - Matrix Representation with Parsimony (MRP) [Baum 92,Ragan 92]
  - Average Consensus Supertree [Lapointe et Cucumel 97]
  - Matrix Representation with Flipping [Chen et al 03, Eulenstein et al. 04]
  - MinCut [Semple et Steel 00, Page 02]

# Plan

I. Problématique – solution
II. Détails de la solution apportée
    1. Obtention des données
    2. Traitement statistique pour montrer des similarités entre distributions de dates
    3. Conversion des similarités de date en information topologique
    4. Utilisation des informations topologiques dans une méthode de super-arbres
III. Synthèse des résultats obtenus
IV. Conclusion

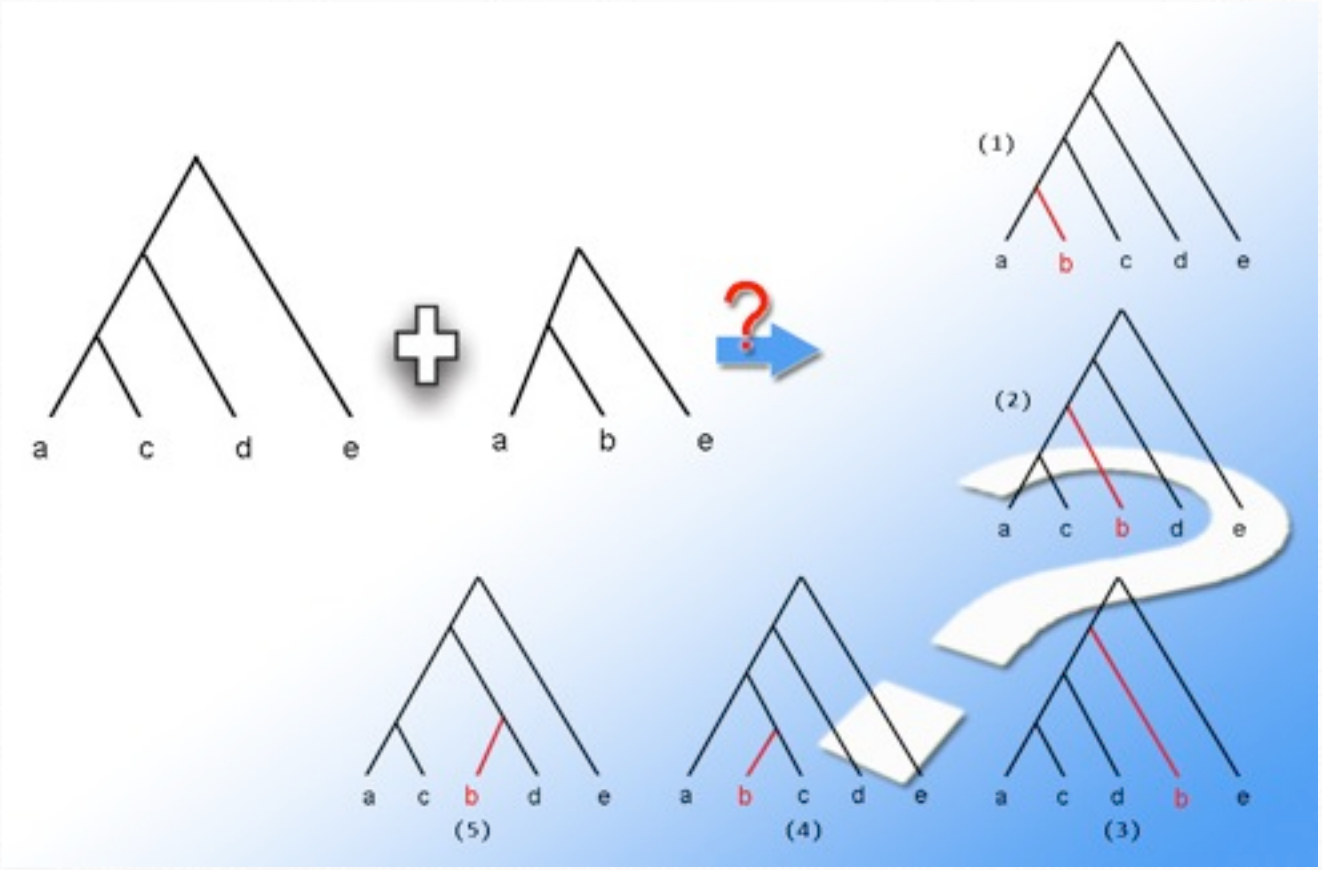
# I. Problématique – solution

# I. Problématique – solution

- Peu de chevauchement entre les ensembles de taxa des phylogénies sources ?

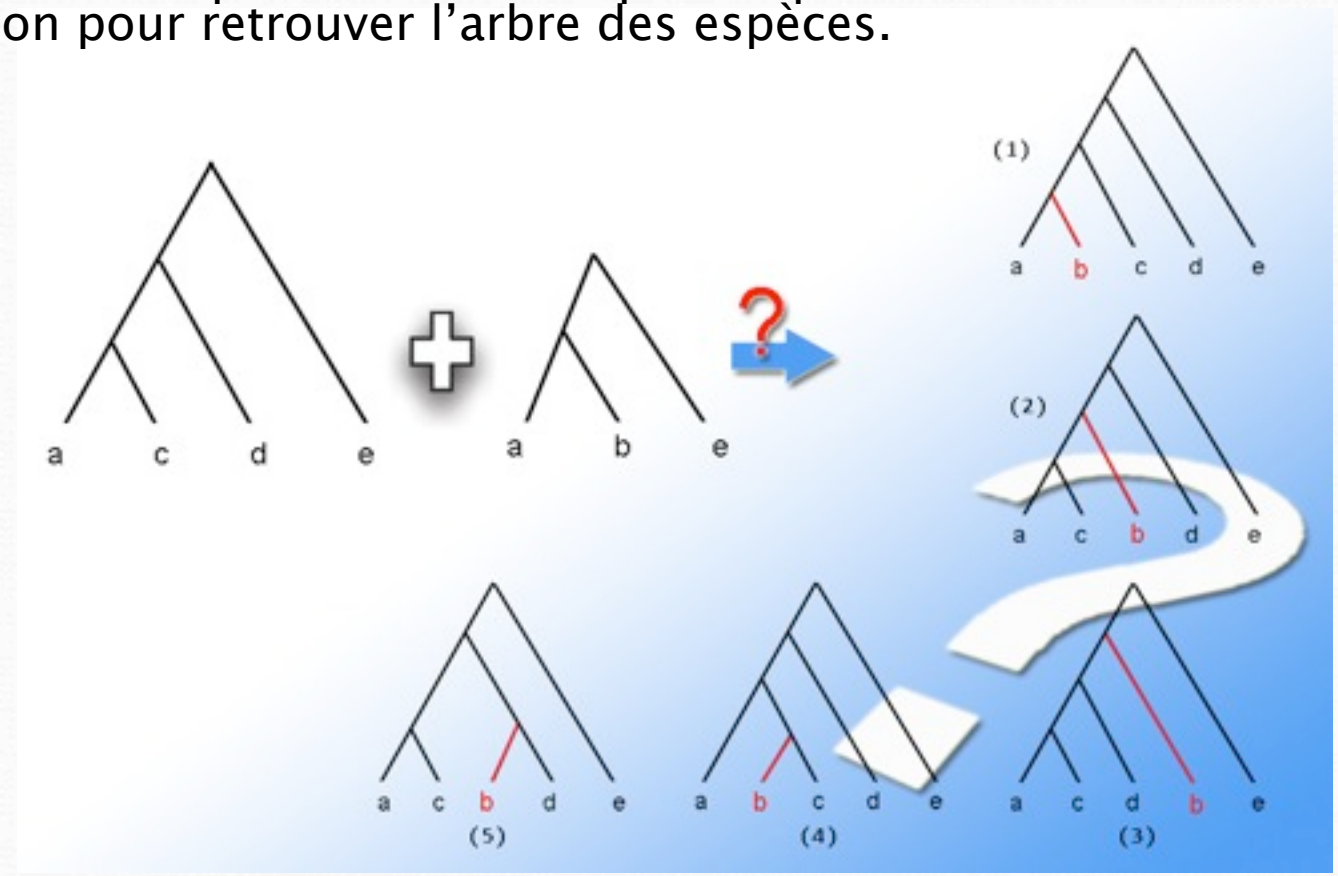
# I. Problématique – solution

- Peu de chevauchement entre les ensembles de taxa des phylogénies sources ?



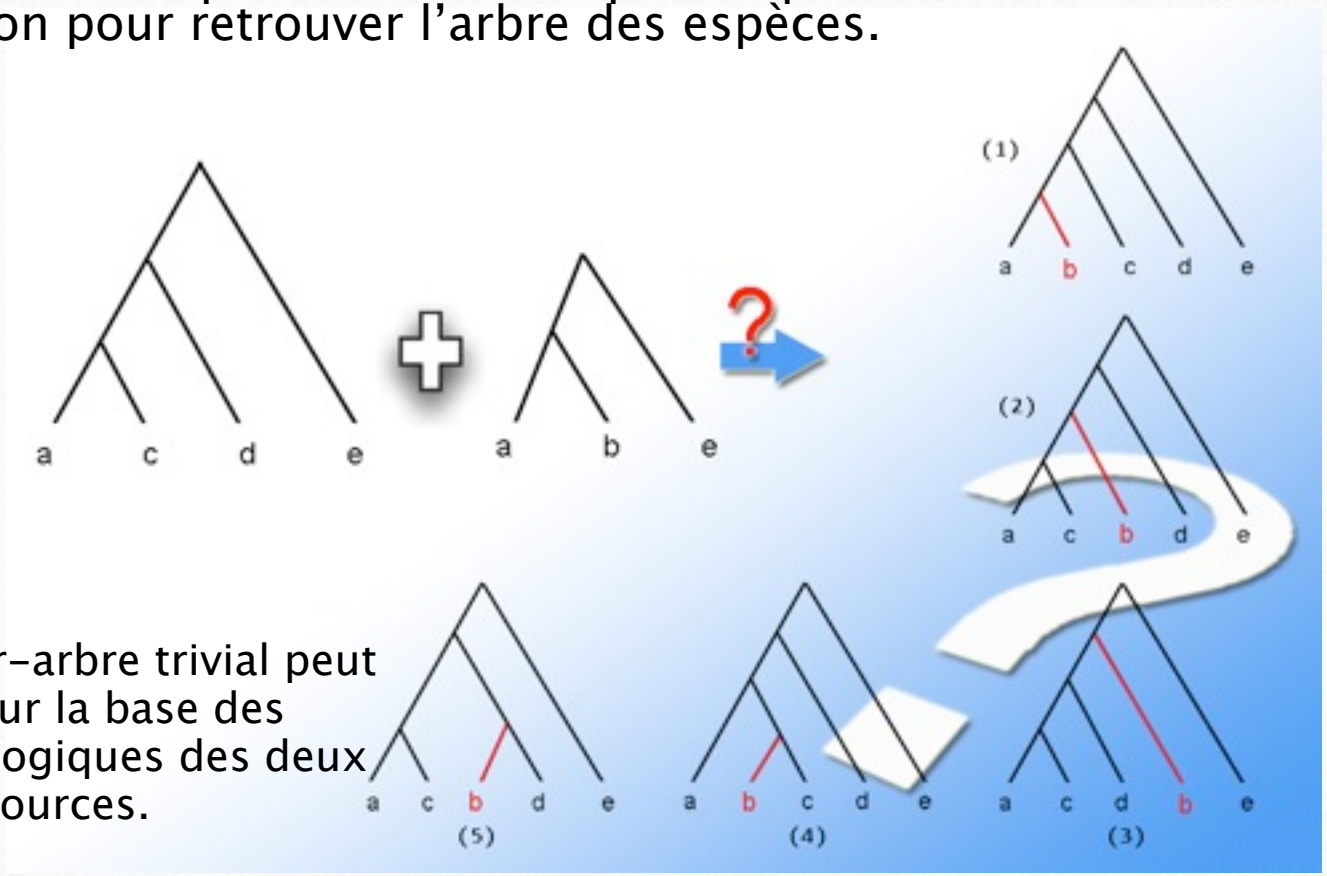
# I. Problématique – solution

- Peu de chevauchement entre les ensembles de taxa des phylogénies sources ?
  - Les méthodes de super-arbres ne disposent pas assez d'information pour retrouver l'arbre des espèces.



# I. Problématique – solution

- Peu de chevauchement entre les ensembles de taxa des phylogénies sources ?
  - Les méthodes de super-arbres ne disposent pas assez d'information pour retrouver l'arbre des espèces.

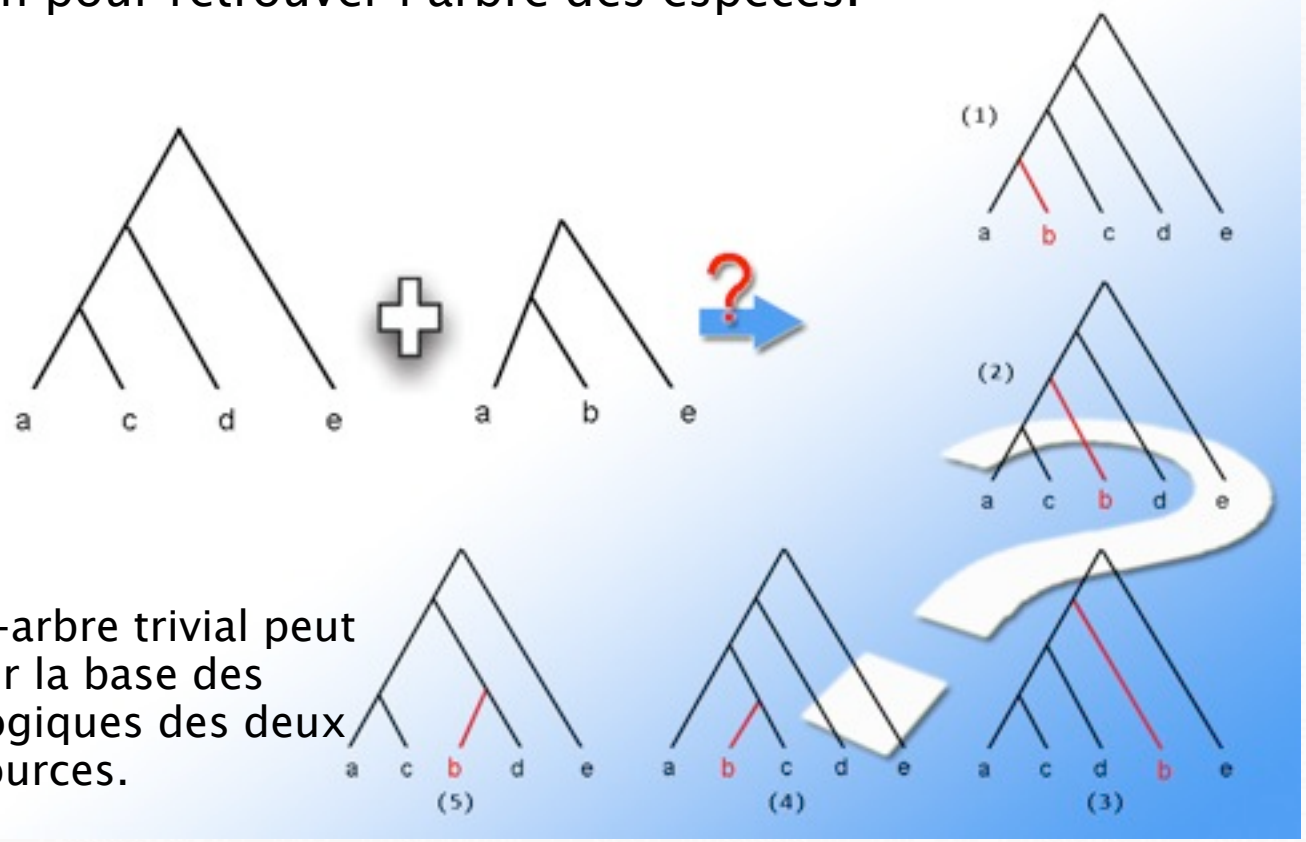



➢ Ici, seul un super-arbre trivial peut être proposé sur la base des informations topologiques des deux arbres sources.

# I. Problématique – solution

- Les méthodes de super-arbres ne disposent pas assez d'information pour retrouver l'arbre des espèces.

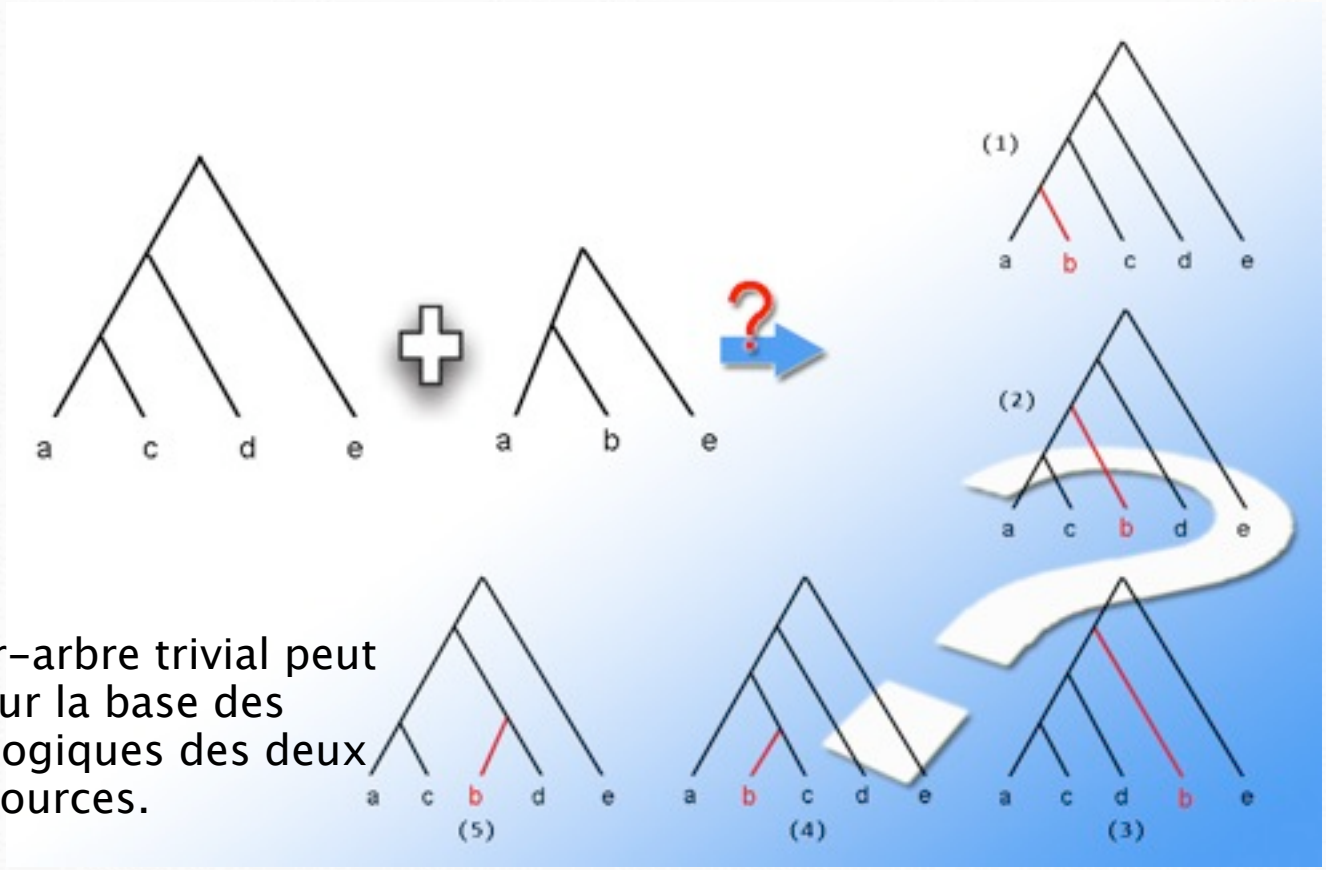
- Solution ?
  - Oui !




➢ Ici, seul un super-arbre trivial peut être proposé sur la base des informations topologiques des deux arbres sources.

# I. Problématique – solution

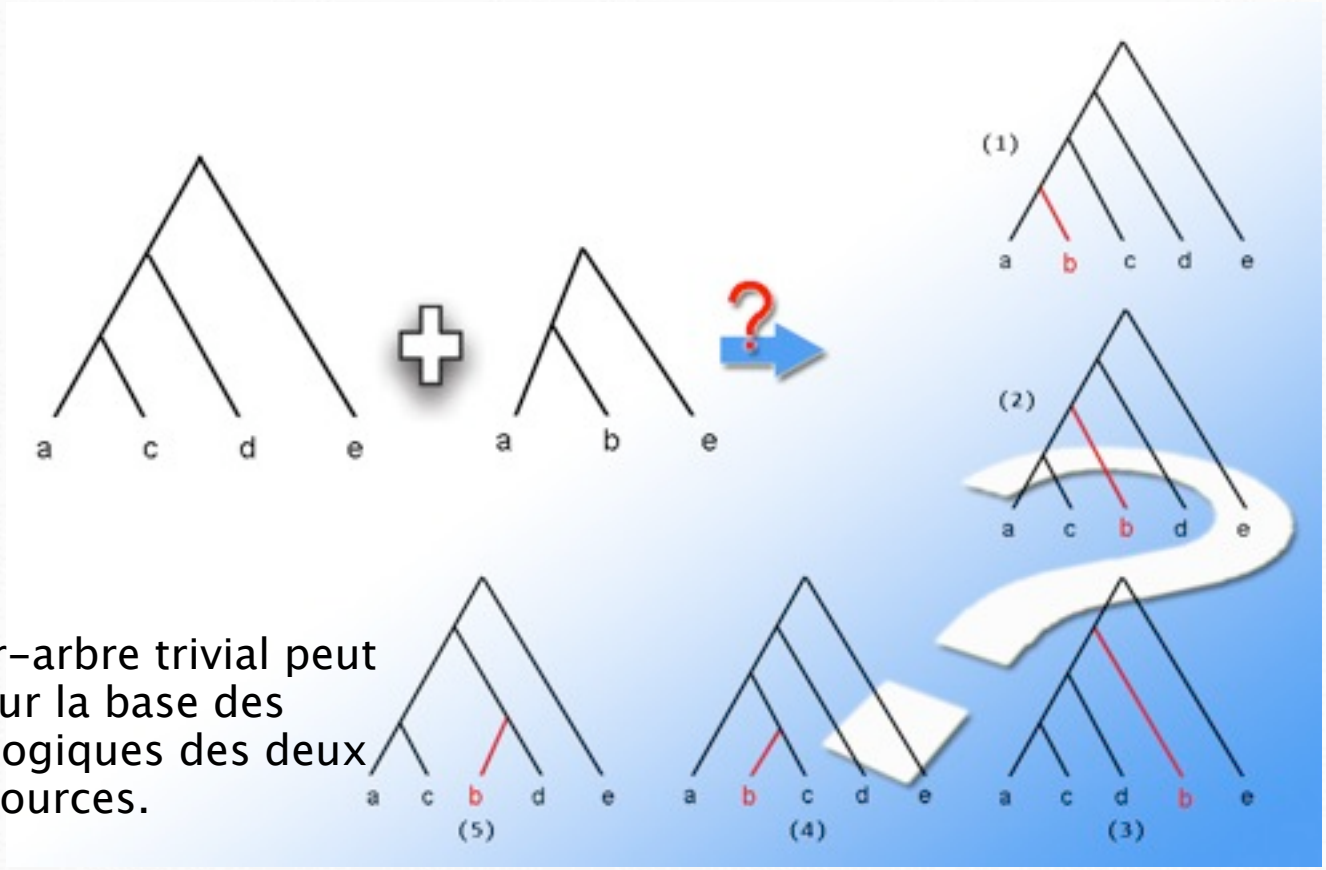
- Solution ?
  - Oui !




➢ Ici, seul un super-arbre trivial peut être proposé sur la base des informations topologiques des deux arbres sources.

# I. Problématique – solution

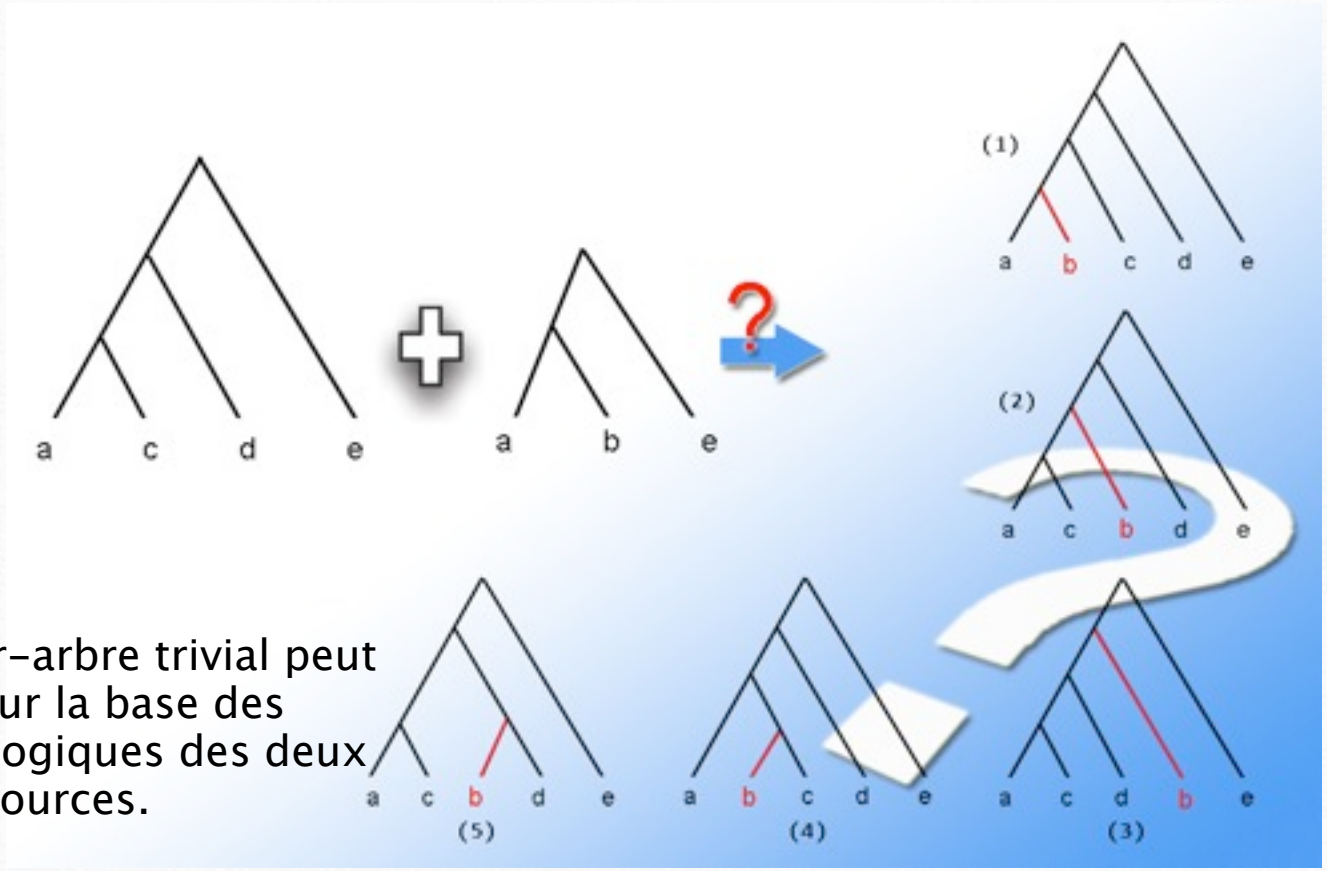
- Solution ?
  - Oui !




➢ Ici, seul un super-arbre trivial peut être proposé sur la base des informations topologiques des deux arbres sources.

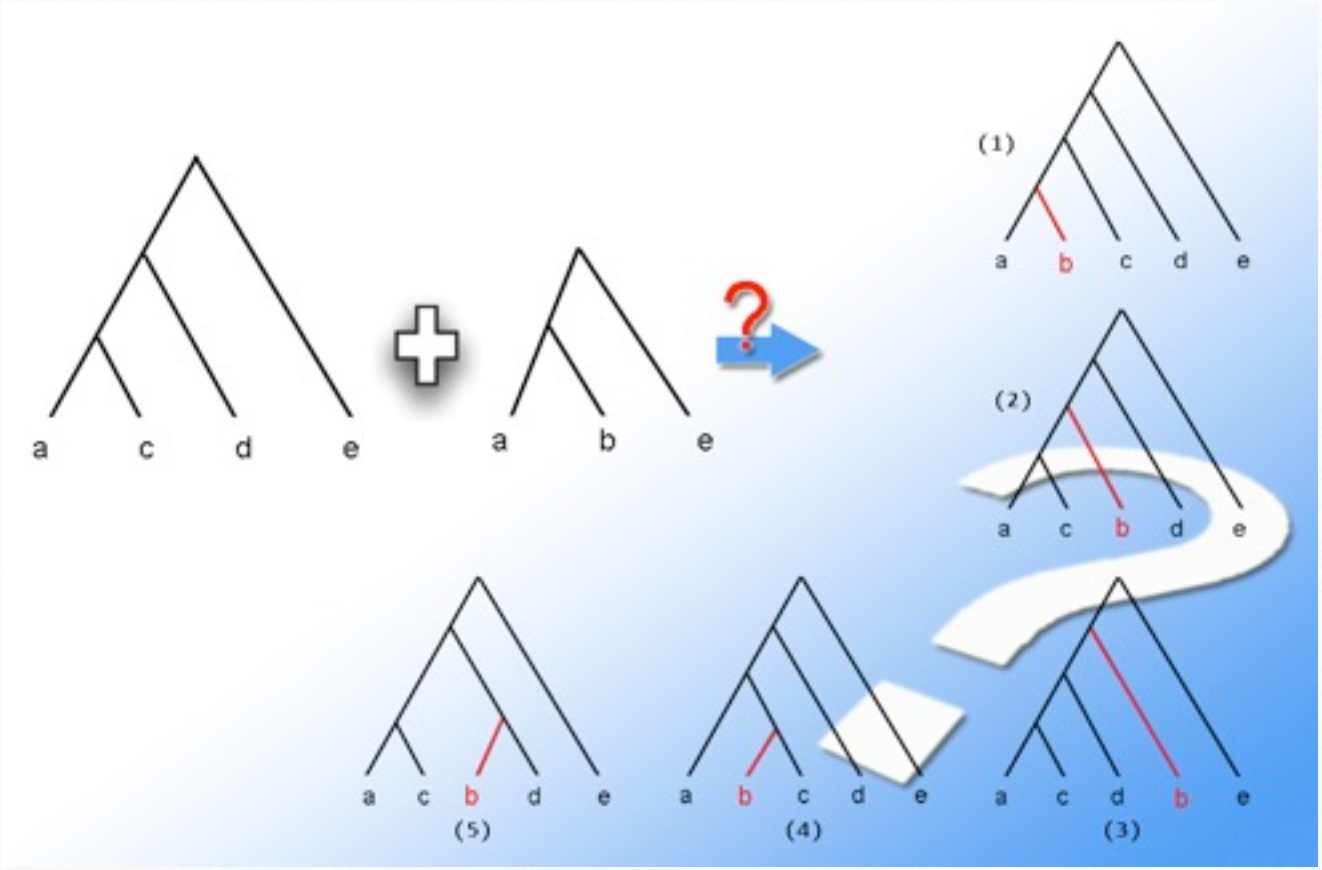
# I. Problématique – solution

- Oui !



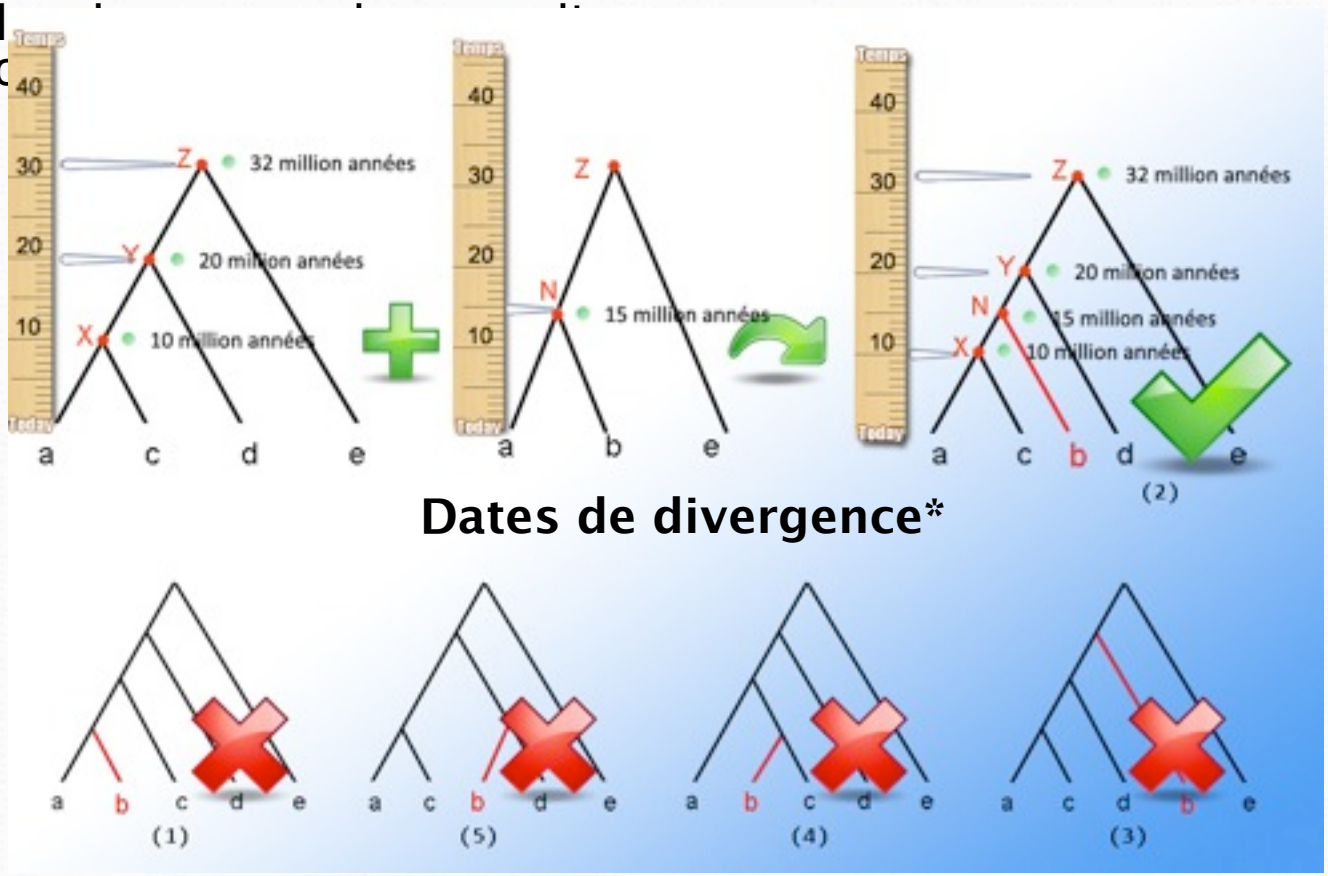

➢ Ici, seul un super-arbre trivial peut être proposé sur la base des informations topologiques des deux arbres sources.

# I. Problématique – solution



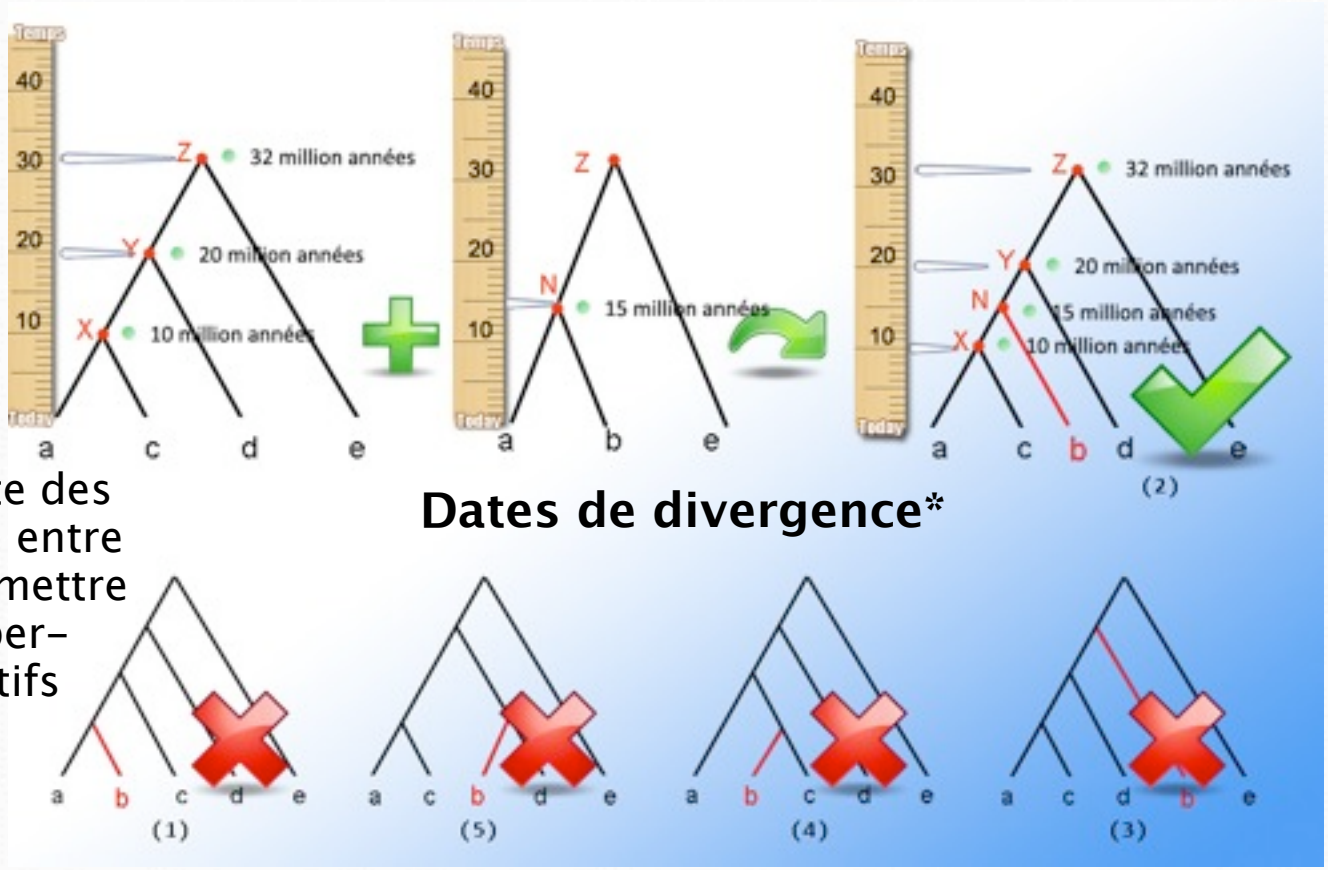
# I. Problématique – solution

- Peu de chevauchement entre les ensembles de taxa des phylogénies sources ?
  - Les méthod
    d'informatic
- Solution ?
  - Oui !




*La **date de divergence** de 2 espèces a et b est l'âge de leur ancêtre commun le plus récent N.

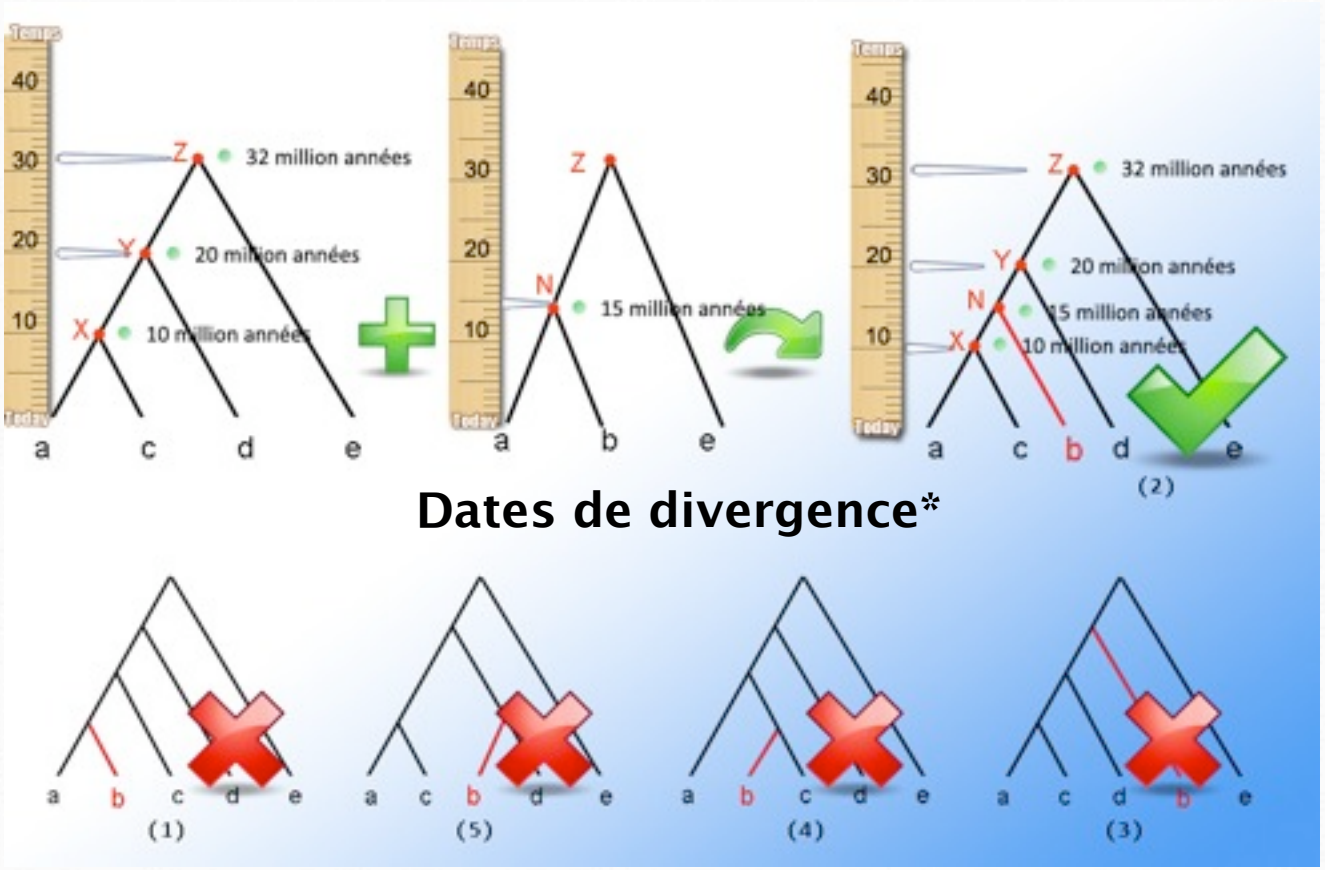
# I. Problématique – solution




- Prendre en compte des dates de divergence entre espèces, il peut permettre de produire des super-arbres plus informatifs

*La **date de divergence** de 2 espèces a et b est l'âge de leur ancêtre commun le plus récent N.

# I. Problématique – solution




*La **date de divergence** de 2 espèces a et b est l'âge de leur ancêtre commun le plus récent N.

# II. Détails de la solution apportée

## 1. Obtention des données

- OrthoMaM 
  - une base de données qui enregistre les marqueurs pour les mammifères (caractérisé par différent paramètres : vitesse d'évolution, nombre d'espèces associées, pourcentage de triplets d'espèces, etc.)
- Un jeu de données très simple
  - Le premier jeu de données vient de la première version de la base de données OrthoMaM (118 exons pour lesquels 12 mammifères sont simultanément présents)
- Un jeu de données plus réaliste
  - basé sur la version actuelle (V4) d'OrthoMaM avec certaines conditions :
    - **CDS (séquence codante)**
    - **10 ≤ species number ≤ 20**
    - **0.9 ≤ relative evolutionary rate ≤ 1.1**
    - **30 ≤ %GC3 ≤ 70**

# II. Détails de la solution apportée

## 1. Obtention des données

- OrthoMaM
  - une base de données qui enregistre les ma mammifères (caractérisé par différent paramè d'évolution, nombre d'espèces associées, pourc d'espèces, etc.)
- Un jeu de données très simple
  - Le premier jeu de données vient de la premiè de données OrthoMaM (118 exons pour lesq sont simultanément présents)
- Un jeu de données plus réaliste
  - basé sur la version actuelle (V4) d'OrthoMaM conditions :
    - **CDS (séquence codante)**
    - **10 ≤ species number ≤ 20**
    - **0.9 ≤ relative evolutionary rate ≤ 1.1**
    - **30 ≤ %GC3 ≤ 70**

# II. Détails de la solution apportée

## 1. Obtention des données

- OrthoMaM 
  - une base de données qui enregistre les marqueurs pour les mammifères (caractérisé par différent paramètres : vitesse d'évolution, nombre d'espèces associées, pourcentage de triplets d'espèces, etc.)
- Un jeu de données très simple
  - Le premier jeu de données vient de la première version de la base de données OrthoMaM (118 exons pour lesquels 12 mammifères sont simultanément présents)
- Un jeu de données plus réaliste
  - basé sur la version actuelle (V4) d'OrthoMaM avec certaines conditions :
    - **CDS (séquence codante)**
    - **10 ≤ species number ≤ 20**
    - **0.9 ≤ relative evolutionary rate ≤ 1.1**
    - **30 ≤ %GC3 ≤ 70**

# 1. Obtention des données

# 1. Obtention des données

- Filtration par PhyloExplorer [Ranwez V.et al. 2009] pour au moins un représentant pour chacun de ces cinq groupes

  La requête de filtre indiquée au site est :

  **{EUARCHONTOGLIRES}>0 and {LAURASIATHERIA}>0 and {AFROTHERIA}>0 and {XENARTHRA}>0 and {MARSUPIALIA }>0 and {Ornithorhynchus}==0**

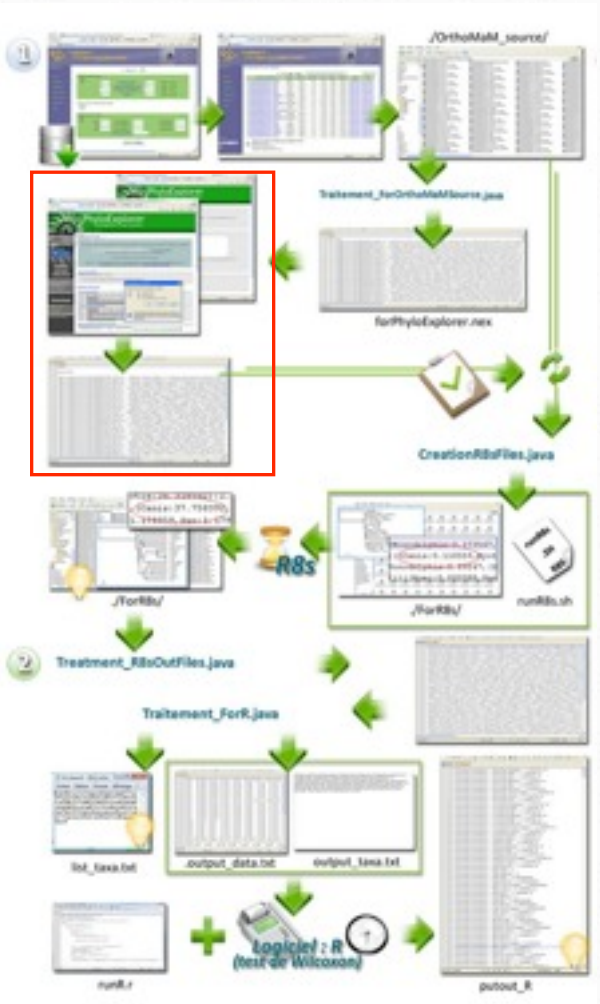
| | | |
|---|---|---|
| **1 EUARCHONTOGLIRES** | | |
| Homo_sapiens | Human | v1 |
| Pan_troglodytes | Chimp | v1 |
| Macaca_mulatta | Macaque | v1 |
| Pongo_pygmaeus | Orangutan | v3 |
| Otolemur_garnettii | Bushbaby | v2 |
| Microcebus_murinus | Mouse lemur | v2 |
| Tupaia_belangeri | Tree shrew | v2 |
| Mus_musculus | Mouse | v1 |
| Rattus_norvegicus | Rat | v1 |
| Cavia_porcellus | Guinea pig | v2 |
| Spermophilus_tridecemlineatus | Ground squirrel | v2 |
| Oryctolagus_cuniculus | Rabbit | v1 |
| Ochotona_princeps | Pika | v2 |
| **2 LAURASIATHERIA** | | |
| Bos_taurus | Cow | v1 |
| Equus_caballus | Horse | v3 |
| Canis_familiaris | Dog | v1 |
| Felis_catus | Cat | v2 |
| Myotis_lucifugus | Microbat | v2 |
| Erinaceus_europaeus | Hedgehog | v2 |
| Sorex_araneus | Shrew | v2 |
| **3 AFROTHERIA** | | |
| Loxodonta_africana | Elephant | v1 |
| Echinops_telfairi | Tenrec | v1 |
| **4 XENARTHRA** | | |
| Dasypus_novemcinctus | Armadillo | v1 |
| **5 MARSUPIALIA** | | |
| Monodelphis_domestica | Opossum | v1 |
| **MONOTREMATA** | | |
| Ornithorhynchus_anatinus | Platypus | v2 |
| | | |

# 1. Obtention des données

- Filtration par PhyloExplorer [Ranwez V.et al. 2009] pour au moins un représentant pour chacun de ces cinq groupes

  La requête de filtre indiquée au site est :

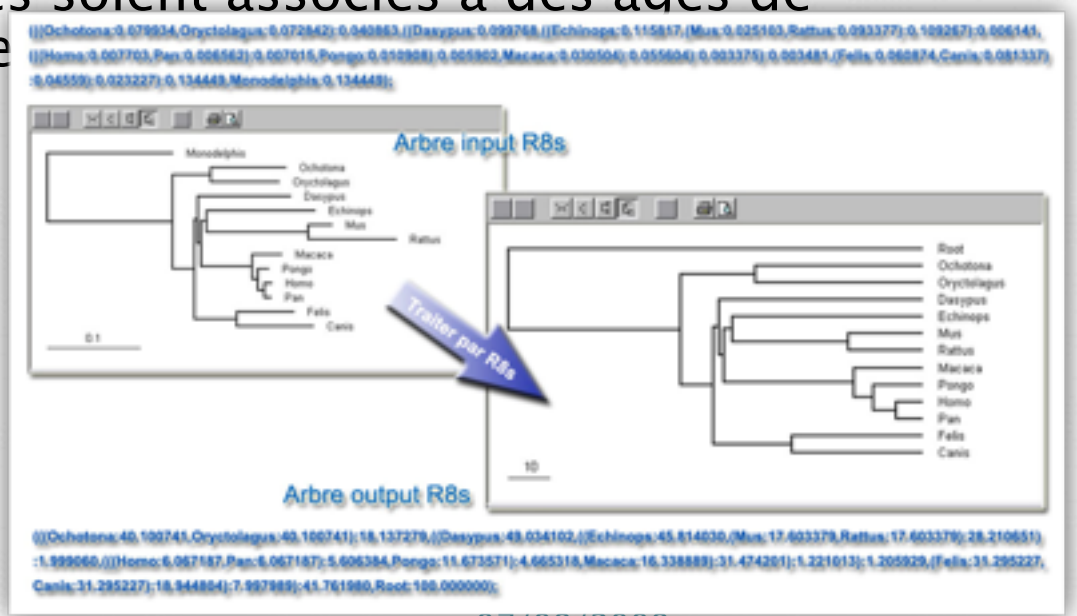
  **{EUARCHONTOGLIRES}>0 and {LAURASIATHERIA}>0 and {AFROTHERIA}>0 and {XENARTHRA}>0 and {MARSUPIALIA }>0 and {Ornithorhynchus}==0**



# 1. Obtention des données

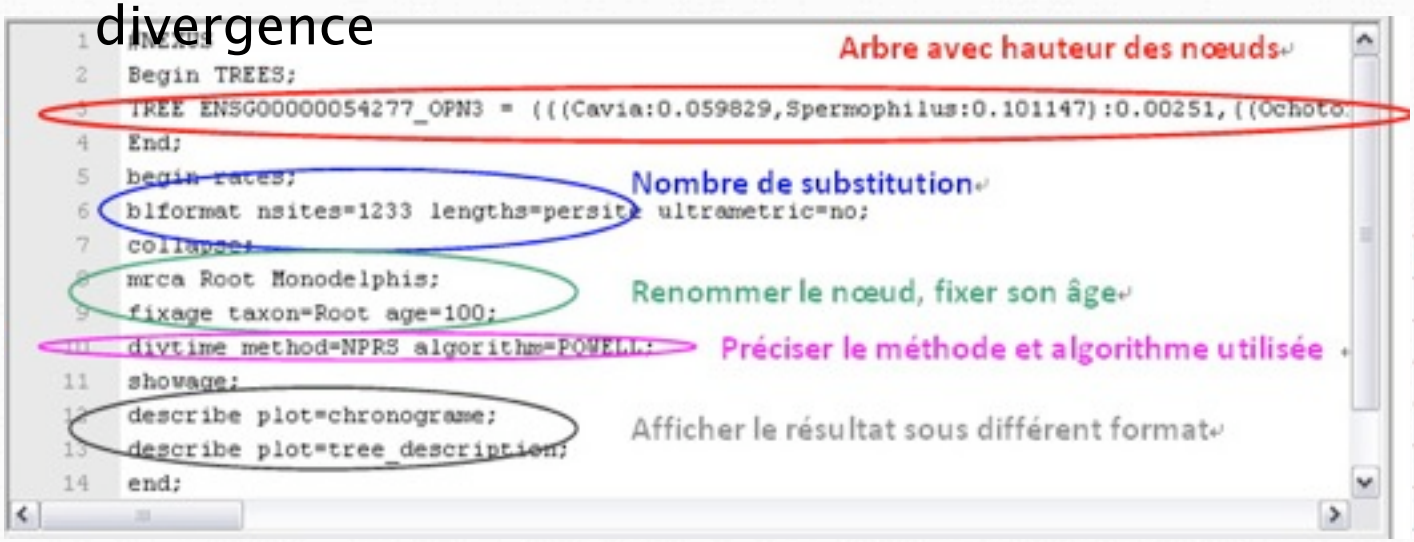
- Filtration par PhyloExplorer [Ranwez V.et al. 2009] pour au moins un représentant pour chacun de ces cinq groupes

- Conversion par R8s [Sanderson, M. J. 2003] pour rendre les arbres ultramétriques et faire que les nœuds internes soient associés à des âges de diverge



# 1. Obtention des données

- Filtration par PhyloExplorer [Ranwez V.et al. 2009] pour au moins un représentant pour chacun de ces cinq groupes
- Conversion par R8s [Sanderson, M. J. 2003] pour rendre les arbres ultramétriques et faire que les nœuds internes soient associés à des âges de divergence




| METHODE | ALGORITHME |
|---|---|
| NPRS | POWELL |
| LF | POWELL |
| LF | TN |
| PL | TN |
| LF | QNEWT |
| PL | QNEWT |

# 2. Traitement statistique pour détecter des similarités entre distributions de dates

# 2. Traitement statistique pour détecter des similarités entre distributions de dates

- Relever toutes les dates de divergence de chaque couple d'espèces
  - Pour chaque couple de taxa, on obtient un ensemble d'estimations de la date du plus récent ancêtre commun de ces 2 taxa

    Par exemple (homo mus : 79.8 78.2 80.1 81.1 … …)

# 2. Traitement statistique pour détecter des similarités entre distributions de dates

- Relever toutes les dates de divergence de chaque couple d'espèces
  - Pour chaque couple de taxa, on obtient un ensemble d'estimations de la date du plus récent ancêtre commun de ces 2 taxa

    Par exemple (homo mus : 79.8 78.2 80.1 81.1 … …)

- Utilisation du logiciel R pour déterminer les similarités existant entre chaque couple d'espèces associé à différentes dates de divergence

# 2. Traitement statistique pour détecter des similarités entre distributions de dates

- Relever toutes les dates de divergence de chaque couple d'espèces
  - Pour chaque couple de taxa, on obtient un ensemble d'estimations de la date du plus récent ancêtre commun de ces 2 taxa
    Par exemple (homo mus : 79.8 78.2 80.1 81.1 … …)

- Utilisation du logiciel R pour déterminer les similarités existant entre chaque couple d'espèces associé à différentes dates de divergence

```
1  Echinops_Loxodonta / Echinops_Dasypus / 3.913637e-14
2  Echinops_Loxodonta / Echinops_Homo / 1.130979e-23
3  Echinops_Loxodonta / Echinops_Pan / 1.130979e-23
4  Echinops_Loxodonta / Echinops_Macaca / 1.130979e-23
5  Echinops_Loxodonta / Echinops_Mus / 1.130979e-23
6  Echinops_Loxodonta / Echinops_Rattus / 1.130979e-23
7  Echinops_Loxodonta / Echinops_Oryctolagus / 1.130979e-23
8  Echinops_Loxodonta / Echinops_Canis / 1.130979e-23
9  Echinops_Loxodonta / Echinops_Bos / 1.130979e-23
10 Echinops_Loxodonta / Loxodonta_Dasypus / 3.913637e-14
11 Echinops_Loxodonta / Loxodonta_Homo / 1.130979e-23
12 Echinops_Loxodonta / Loxodonta_Pan / 1.130979e-23
```

# 2. Traitement statistique pour détecter des similarités entre distributions de dates

- Relever toutes les dates de divergence de chaque couple d'espèces
  - Pour chaque couple de taxa, on obtient un ensemble d'estimations de la date du plus récent ancêtre commun de ces 2 taxa

    Par exemple (homo mus : 79.8 78.2 80.1 81.1 … …)

- Utilisation du logiciel R pour déterminer les similarités existant entre chaque couple d'espèces associé à différentes dates de divergence

p-valeur petite : Avoir plus de chance que ces deux distributions ne soient pas issues de la même loi.
p-valeur grande : Ne pas pouvoir dire qu'elles sont différentes, autrement dit on conclut pragmatiquement qu'elles sont identiques

# 2. Traitement statistique pour détecter des similarités entre distributions de dates

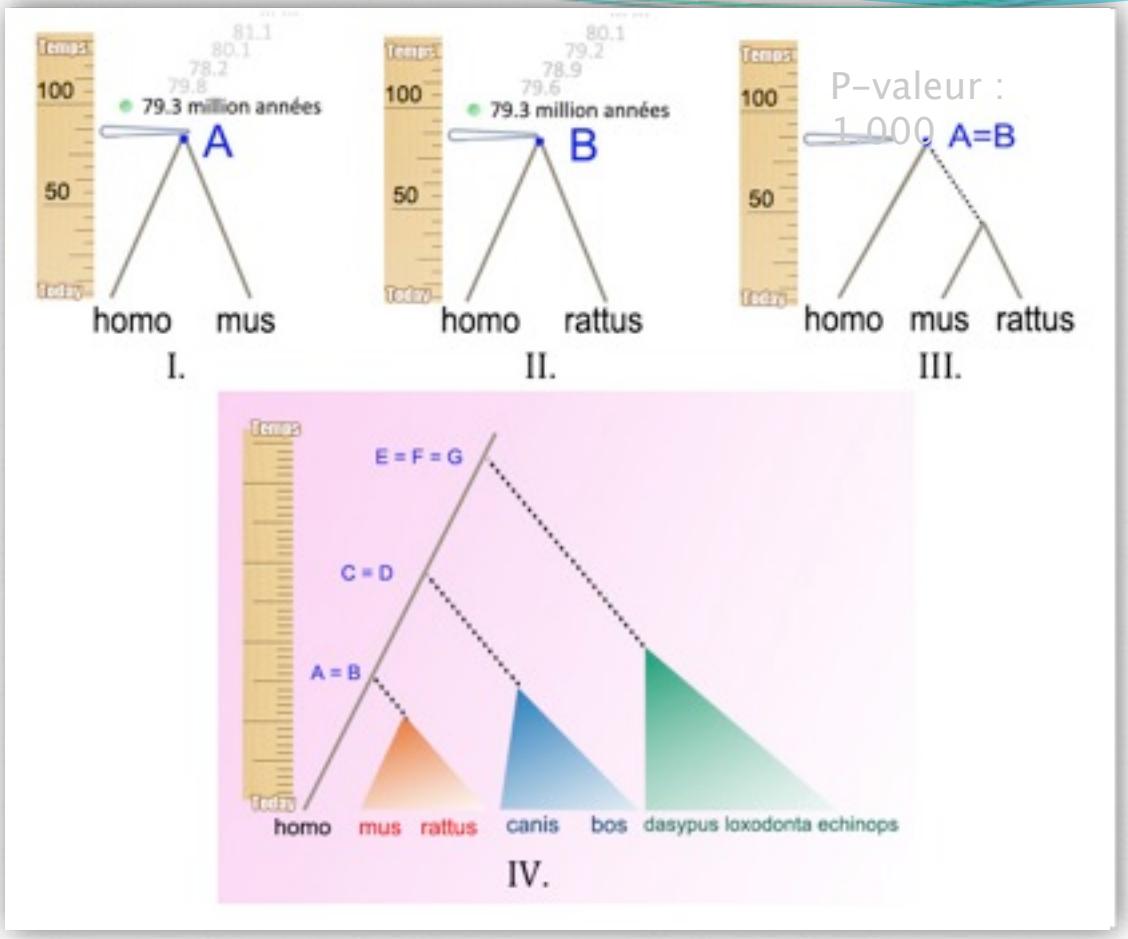
- Relever toutes les dates de divergence de chaque couple d'espèces
  - Pour chaque couple de taxa, on obtient un ensemble d'estimations de la date du plus récent ancêtre commun de ces 2 taxa

    Par exemple (homo mus : 79.8 78.2 80.1 81.1 … …)

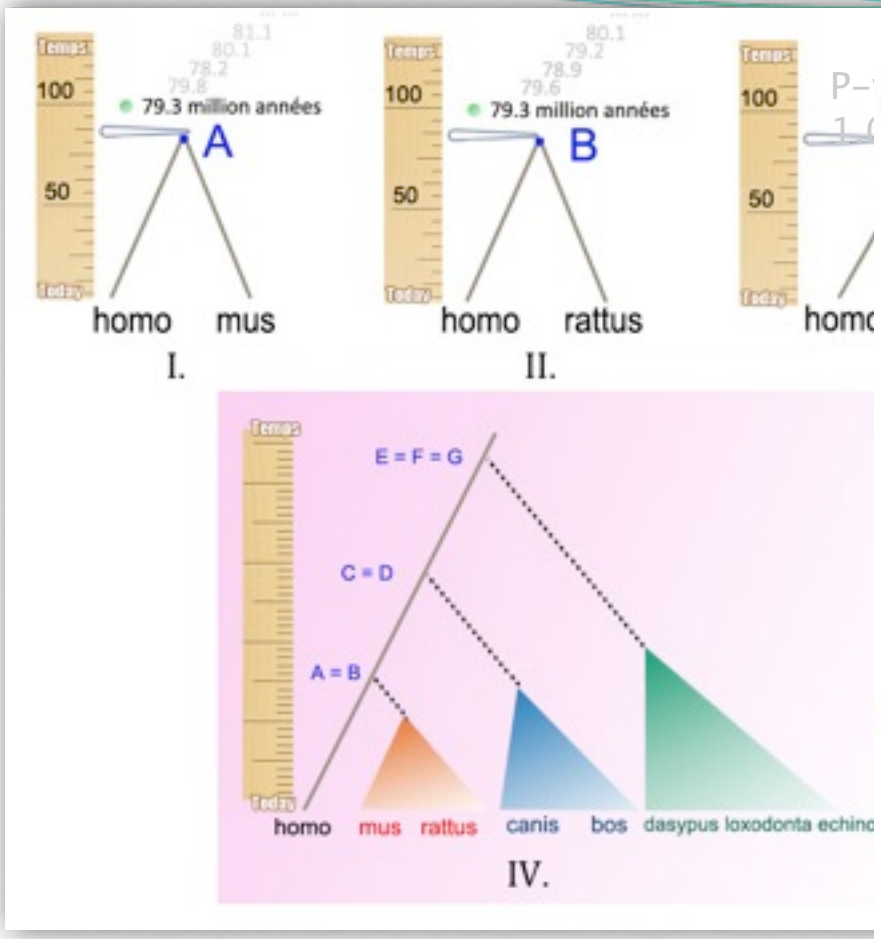
- Utilisation du logiciel R pour déterminer les similarités existant entre chaque couple d'espèces associé à différentes dates de divergence
  - Test de Wilcoxon [Wilcoxon, F. 1945] : permettre de tester si deux échantillons peuvent être issus de la même loi.

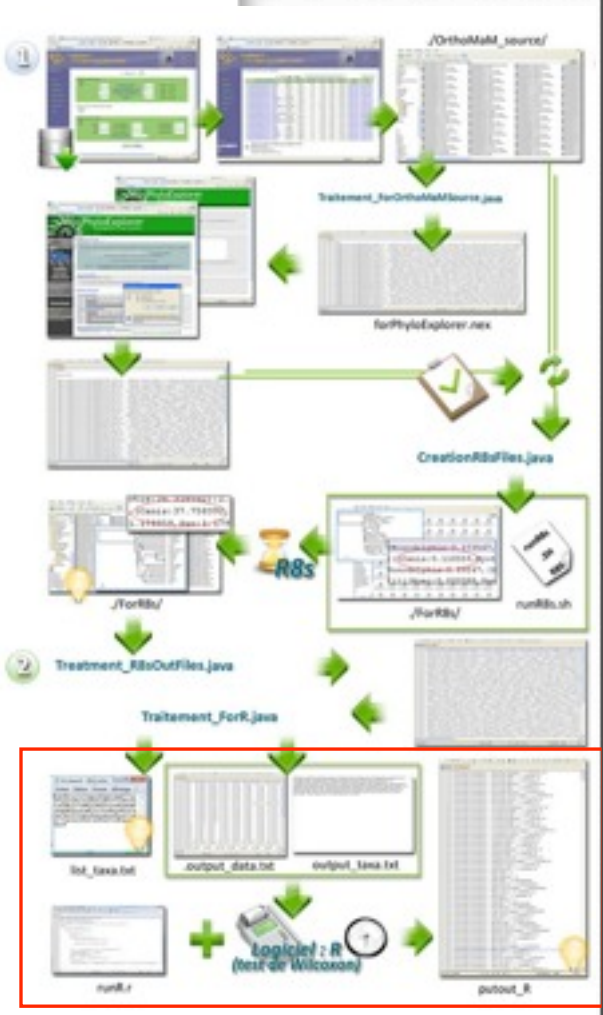
- On s'intéresse à deux couples qui contiennent une espèce en commun

  Par exemple : mus/homo et rattus/homo.

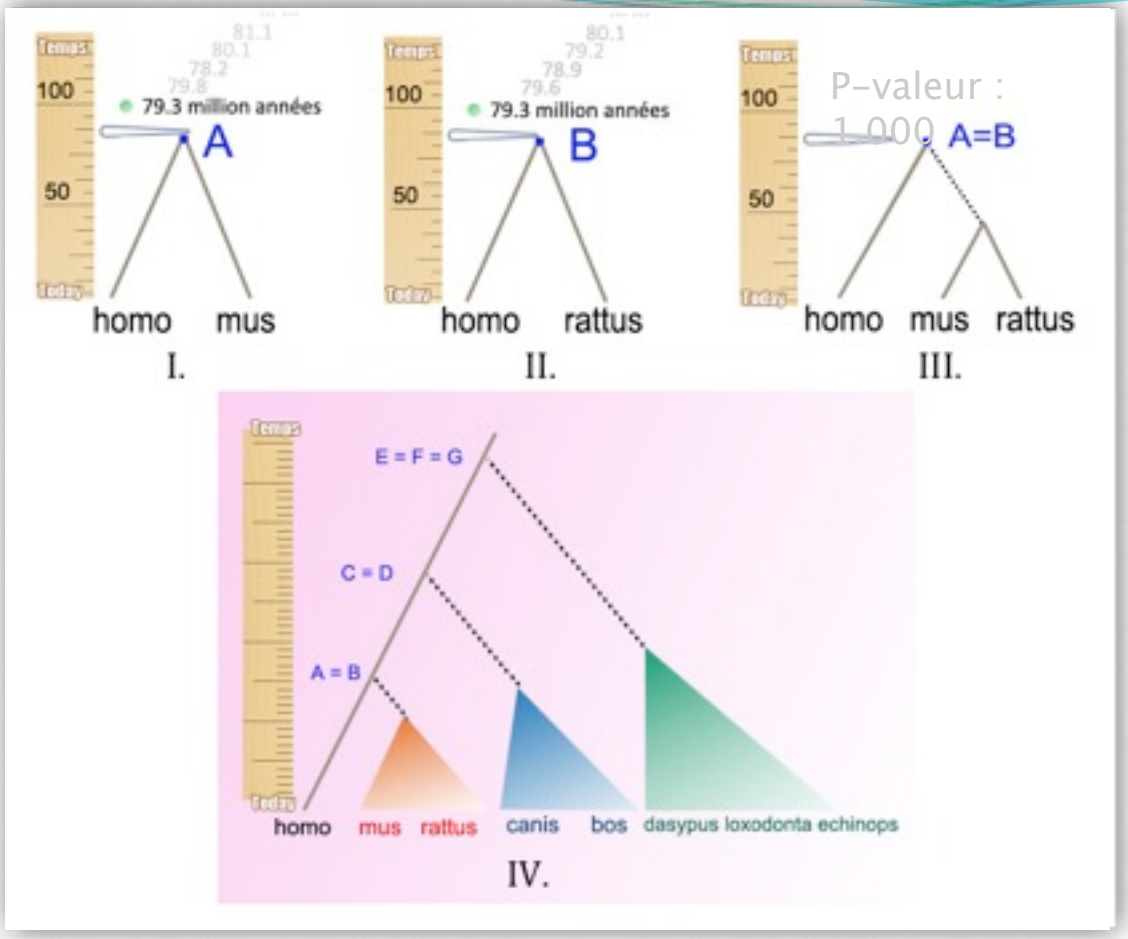



I. Estimations de l'âge de A (ancêtre le plus récent de homo et mus)
II. Estimations de l'âge de B (ancêtre le plus récent de homo et rattus)
III. Le test de Wilcoxon permet ici de conclure que A et B sont en fait la même espèce ancestrale
IV. Objectif : identification de groupes taxonomique en fonction des tests sur tous les couples référençant une même espèce (ici homo)




I. Estimations de l'âge de A (ancêtre le plus récent de
II. Estimations de l'âge de B (ancêtre le plus récent de l
III. Le test de Wilcoxon permet ici de conclure que A et B sont en fait la même espèce ancestrale
IV. Objectif : identification de groupes taxonomique en fonction des tests sur tous les couples référençant une même espèce (ici homo)




I. Estimations de l'âge de A (ancêtre le plus récent de homo et mus)
II. Estimations de l'âge de B (ancêtre le plus récent de homo et rattus)
III. Le test de Wilcoxon permet ici de conclure que A et B sont en fait la même espèce ancestrale
IV. Objectif : identification de groupes taxonomique en fonction des tests sur tous les couples référençant une même espèce (ici homo)

# 3. Conversion des similarités de dates en information topologique

# 3. Conversion des similarités de dates en information topologique

- Créer une présentation de ces informations sous forme de graphes
  - Choisir une espèce x comme référence,
  - Choisir une p-valeur seuil pour interpréter les résultats des tests des couples contenant cette espèce référente.
  - Inclure un sommet pour chaque espèce du jeu de données, sauf x.
  - Si la p-valeur entre x/y et x/z > seuil choisi, on met une arête entre y et z.

# 3. Conversion des similarités de dates en information topologique

- Créer une présentation de ces informations sous forme de graphes
  - Choisir une espèce x comme référence,
  - Choisir une p-valeur seuil pour interpréter les résultats des tests des couples contenant cette espèce référente.
  - Inclure un sommet pour chaque espèce du jeu de données, sauf x.
  - Si la p-valeur entre x/y et x/z > seuil choisi, on met une arête entre y et z.
- Idéalement : représenter un ensemble de cliques*
  - Correspondant aux groupes d'espèces qui partagent un même ancêtre commun avec l'espèce référente x.

*Clique : un ensemble C de sommets tel que pour tous sommets y et z de C, il existe une arête entre y et z.

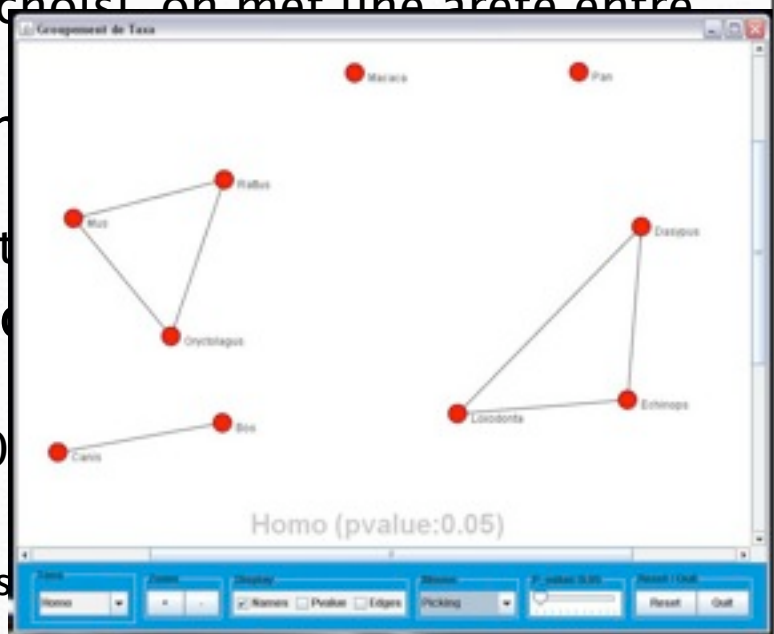
# 3. Conversion des similarités de dates en information topologique

- Créer une présentation de ces informations sous forme de graphes
  - Choisir une espèce x comme référence,
  - Choisir une p-valeur seuil pour interpréter les résultats des tests des couples contenant cette espèce référente.
  - Inclure un sommet pour chaque espèce du jeu de données, sauf x.
  - Si la p-valeur entre x/y et x/z > seuil choisi, on met une arête entre y et z.
- Idéalement : représenter un ensemble de cliques*
  - Correspondant aux groupes d'espèces qui partagent un même ancêtre commun avec l'espèce référente x.
- P-valeur seuil les plus faibles = moins de chance de se tromper
  - Un compromis à trouver : de 0.001 à 0.2

*Clique : un ensemble C de sommets tel que pour tous sommets y et z de C, il existe une arête entre y et z.

# 3. Conversion des similarités de dates en information topologique

- Créer une présentation de ces informations sous forme de graphes
  - Choisir une espèce x comme référence,
  - Choisir une p-valeur seuil pour interpréter les résultats des tests des couples contenant cette espèce référente.
  - Inclure un sommet pour chaque espèce du jeu de données, sauf x.
  - Si la p-valeur entre x/y et x/z > seuil choisi, on met une arête entre y et z.
- Idéalement : représenter un ensem
  - Correspondant aux groupes d'espèces ancêtre commun avec l'espèce référent
- P-valeur seuil les plus faibles = m tromper
  - Un compromis à trouver : de 0.001 à 0

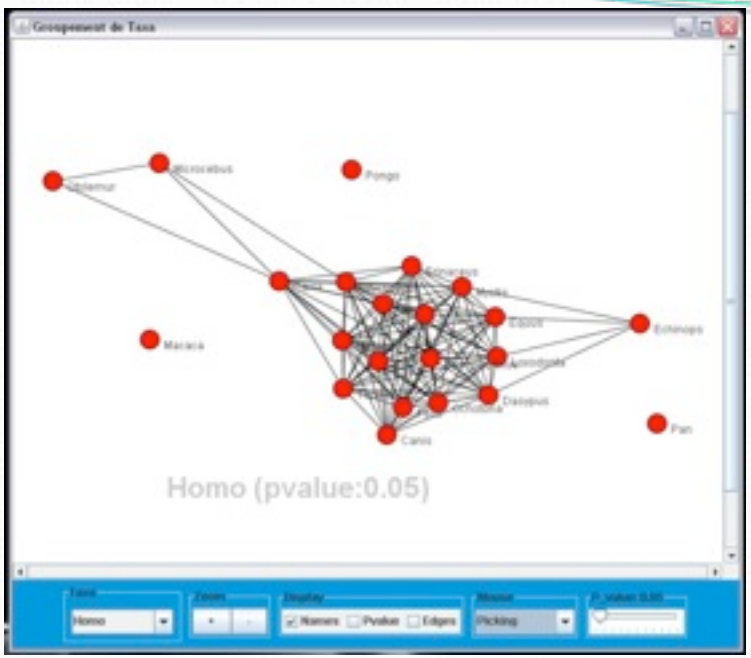



*Clique : un ensemble C de sommets tel que pour tous s arête entre y et z.

# 3. Conversion des similarités de dates en information topologique

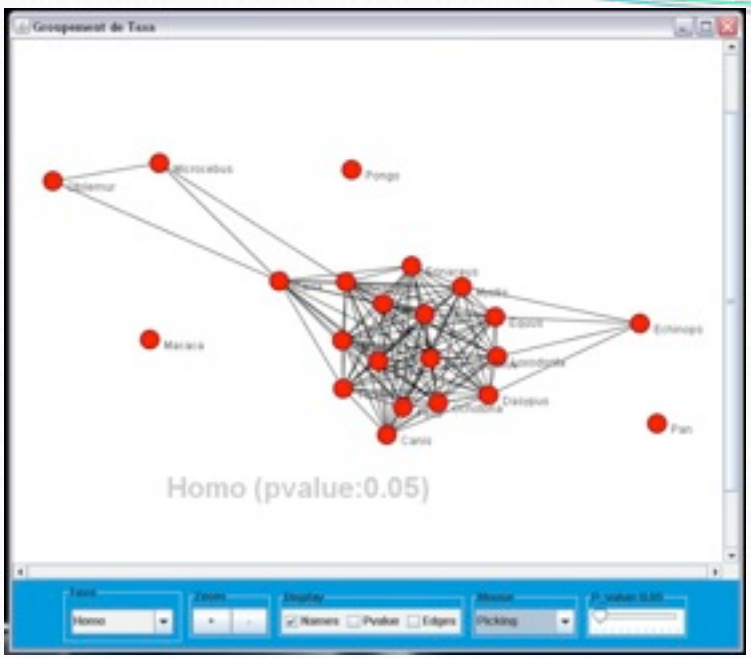
- Créer une présentation de ces informations sous forme de graphes
  - Choisir une espèce x comme référence,
  - Choisir une p-valeur seuil pour interpréter les résultats des tests des couples contenant cette espèce référente.
  - Inclure un sommet pour chaque espèce du jeu de données, sauf x.
  - Si la p-valeur entre x/y et x/z > seuil choisi, on met une arête entre y et z.
- Idéalement : représenter un ensemble de cliques*
  - Correspondant aux groupes d'espèces qui partagent un même ancêtre commun avec l'espèce référente x.
- P-valeur seuil les plus faibles = moins de chance de se tromper
  - Un compromis à trouver : de 0.001 à 0.2
- Convertir les similarités en information topologique (un arbre) :
  - ((Pan, Macaca, (Rattus, Mus, Oryctolagus), (Canis, Bos), (Echinops,

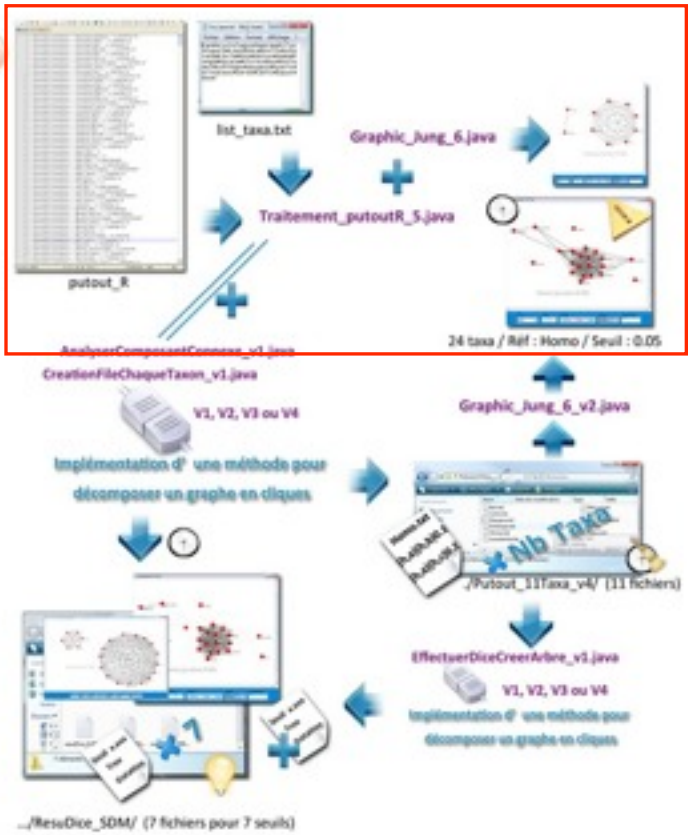
*Clique : un ensemble C de sommets tel que pour tous sommets y et z de C, il existe une arête entre y et z.




(Deuxième jeu de données / Réf taxon: Homo / Seuil p-valeur : 0.05)

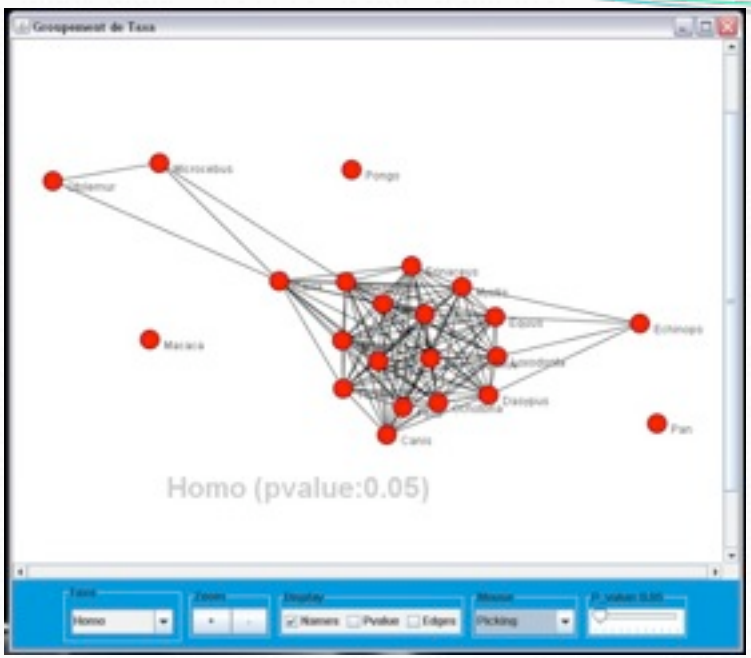
- Pour le deuxième jeu de données
  - On n'a pas toujours des cliques
  - Choisir un petit seuil pour avoir plus de fiabilité (La plupart des arêtes sont encore présentes.)
  - Nécessaire pour transformer ce graphe en une union de cliques disjointes




(Deuxième jeu de données / Réf taxon: Hom

- Pour le deuxième jeu de données
  - On n'a pas toujours des cliques
  - Choisir un petit seuil pour avoir plus de flabilite (La plupart des arêtes sont encore présentes.)
  - Nécessaire pour transformer ce graphe en une union de cliques disjointes




(Deuxième jeu de données / Réf taxon: Homo / Seuil p-valeur : 0.05)

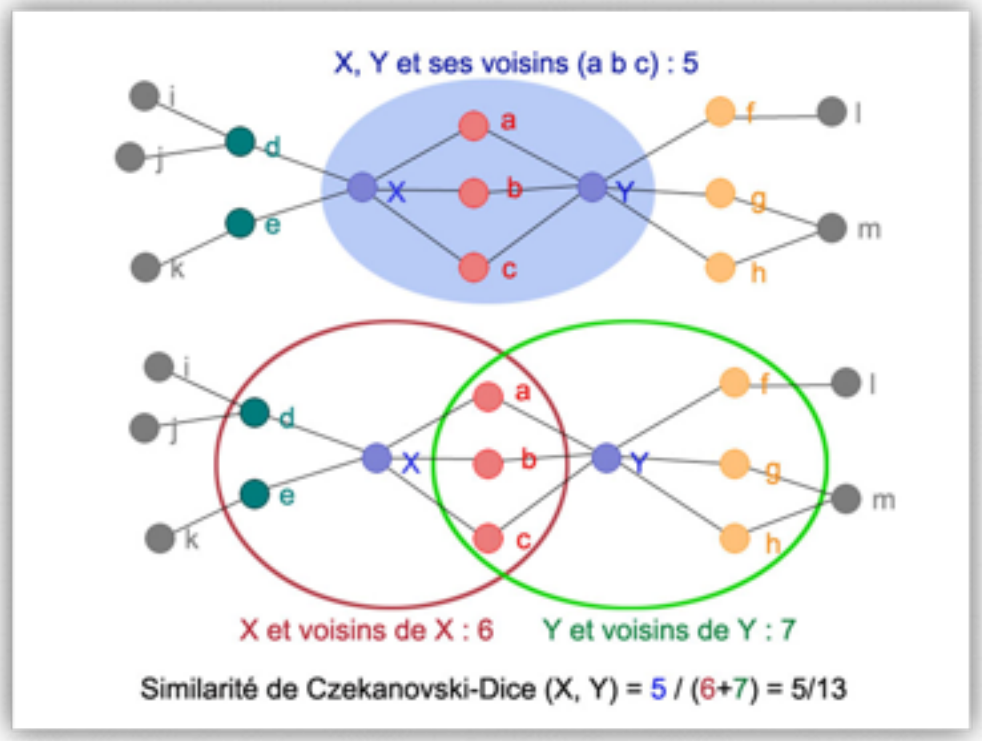
- Pour le deuxième jeu de données
  - On n'a pas toujours des cliques
  - Choisir un petit seuil pour avoir plus de fiabilité (La plupart des arêtes sont encore présentes.)
  - Nécessaire pour transformer ce graphe en une union de cliques disjointes

- Implémentation d'une méthode pour décomposer un graphe en cliques

  1. Méthode similarité de Czekanovski-Dice [Dice 1945] V1

- Implémentation d'une méthode pour décomposer un graphe en cliques

1. Méthode similarité de Czekanovski-Dice [Dice 1945] V1



- Implémentation d'une méthode pour décomposer un graphe en cliques

1. Méthode similarité de Czekanovski-Dice [Dice 1945] V1
   - En fonction de leurs sommets voisins partagés et non partagés
   - Prendre en compte la simple présence ou absence d'arêtes entre deux sommets

2. Méthode similarité de Czekanovski-Dice optimal V2
   - Quand il y a plusieurs arêtes avec la plus faible similarité on prendre en compte le poids d'arête
   - Enlever l'arête qui possède la p-valeur la plus faible

- Implémentation d'une méthode pour décomposer un graphe en cliques

1. Méthode similarité de Czekanovski-Dice [Dice 1945] V1
   - En fonction de leurs sommets voisins partagés et non partagés
   - Prendre en compte la simple présence ou absence d'arêtes entre deux sommets

2. Méthode similarité de Czekanovski-Dice optimal V2
   - Quand il y a plusieurs arêtes avec la plus faible similarité on prendre en compte le poids d'arête
   - Enlever l'arête qui possède la p-valeur la plus faible

- Implémentation d'une méthode pour décomposer un graphe en cliques

1. Méthode similarité de Czekanovski-Dice [Dice 1945] V1
   - En fonction de leurs sommets voisins partagés et non partagés
   - Prendre en compte la simple présence ou absence d'arêtes entre deux sommets

2. Méthode similarité de Czekanovski-Dice optimal V2
   - Quand il y a plusieurs arêtes avec la plus faible similarité on prendre en compte le poids d'arête
   - Enlever l'arête qui possède la p-valeur la plus faible

3. Méthode quand les pondérations sont des intensités [Angelelli et al. 2008] V3

- Implémentation d'une méthode pour décomposer un graphe en cliques

1. Méthode similarité de Czekanovski-Dice [Dice 1945] V1
   - En fonction de leurs sommets voisins partagés et non partagés
   - Prendre en compte la simple présence ou absence d'arêtes entre deux sommets

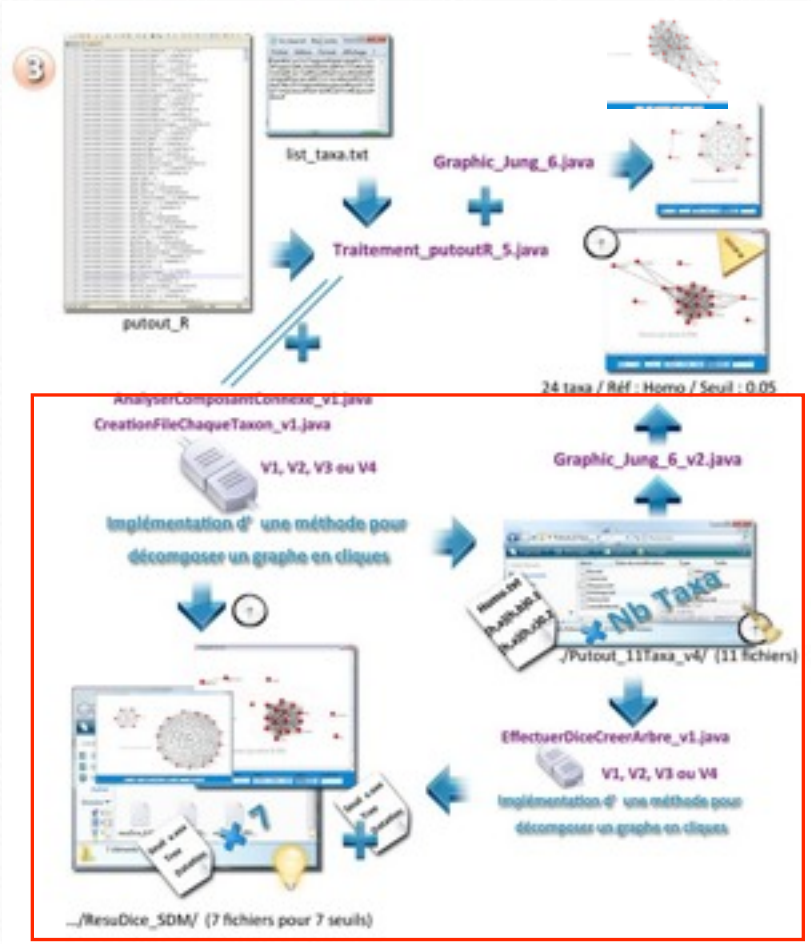
2. Méthode similarité de Czekanovski-Dice optimal V2
   - Quand il y a plusieurs arêtes avec la plus faible similarité on prendre en compte le poids d'arête
   - Enlever l'arête qui possède la p-valeur la plus faible

3. Méthode quand les pondérations sont des intensités [Angelelli et al. 2008] V3

- Effectuer ces quatre méthodes pour les différentes p-valeurs seuils et chaque espèce prise comme référence

  Seuil : 0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2

- Création des fichiers qui contiennent les couples d'espèces et les p-valeurs pour chaque espèce prise comme référence
  - Créer la deuxième version du programme pour présenter les similarités des espèces graphiquement et surtout efficacement
  - Créer une classe pour convertir



# 4. Utilisation des informations topologiques dans une méthode de super-arbres

- Fichiers obtenus : une collection d'arbres par groupement des espèces qui sont issu d'une même date de divergence
- Ajouter les arbres source pour construire une nouvelle collection des arbres totaux. Enfin, on va construire un super-arbre avec les deux méthodes :
  - Méthode normale : utiliser les arbres source, il contient que les informations de topologie
  - Nouvelle méthode : utiliser les arbres source + arbres obtenus par les dates de divergence
- Par rapport à la méthode normale, on va tester notre nouvelle méthode pour savoir si la fiabilité du résultat ou l'information obtenue est plus grande

## ➢ Obtention des super-arbres

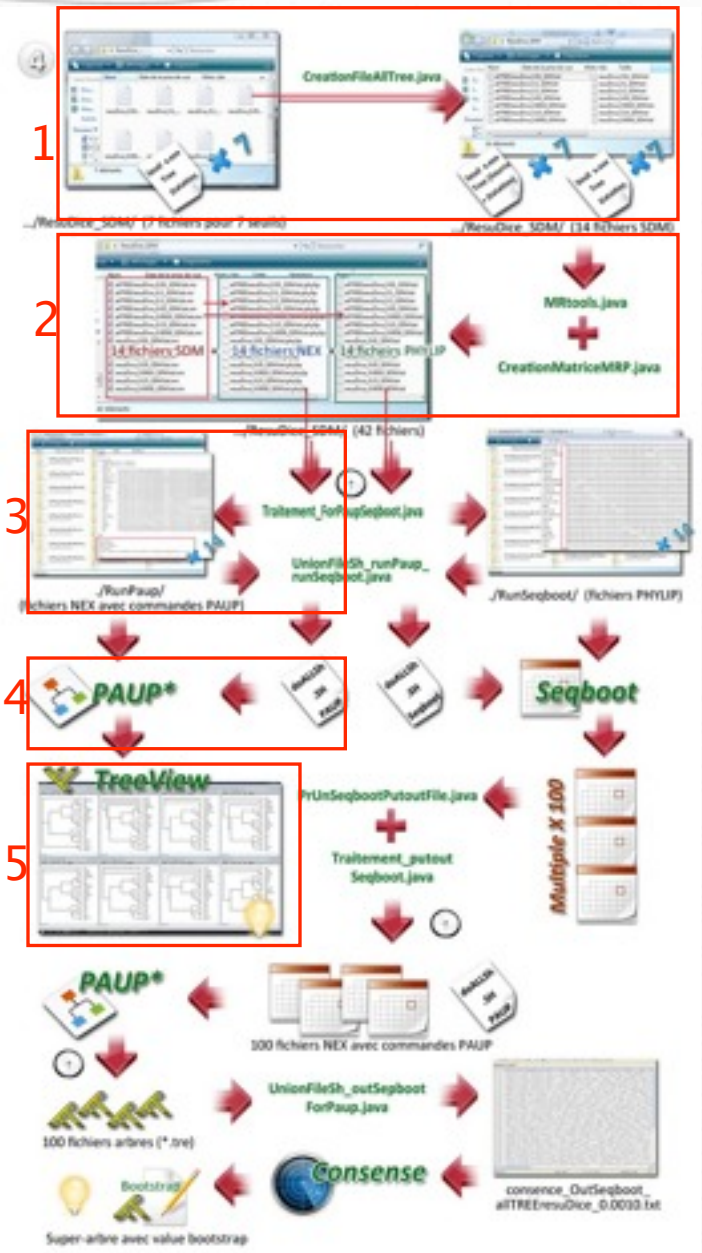
1. Combiner les arbres source et les arbres qui ont été obtenu par les dates de divergence dans un même fichier.

2. Créer un fichier contenant une représentation matricielle binaire de la collection d'arbres sources par MRtools [Alexis Criscoulo 2007]

3. Ajouter les commandes PAUP pour chaque fichier d'output MRtools (Format NEXUS) et crée un fichier d'exécution runPaup.sh

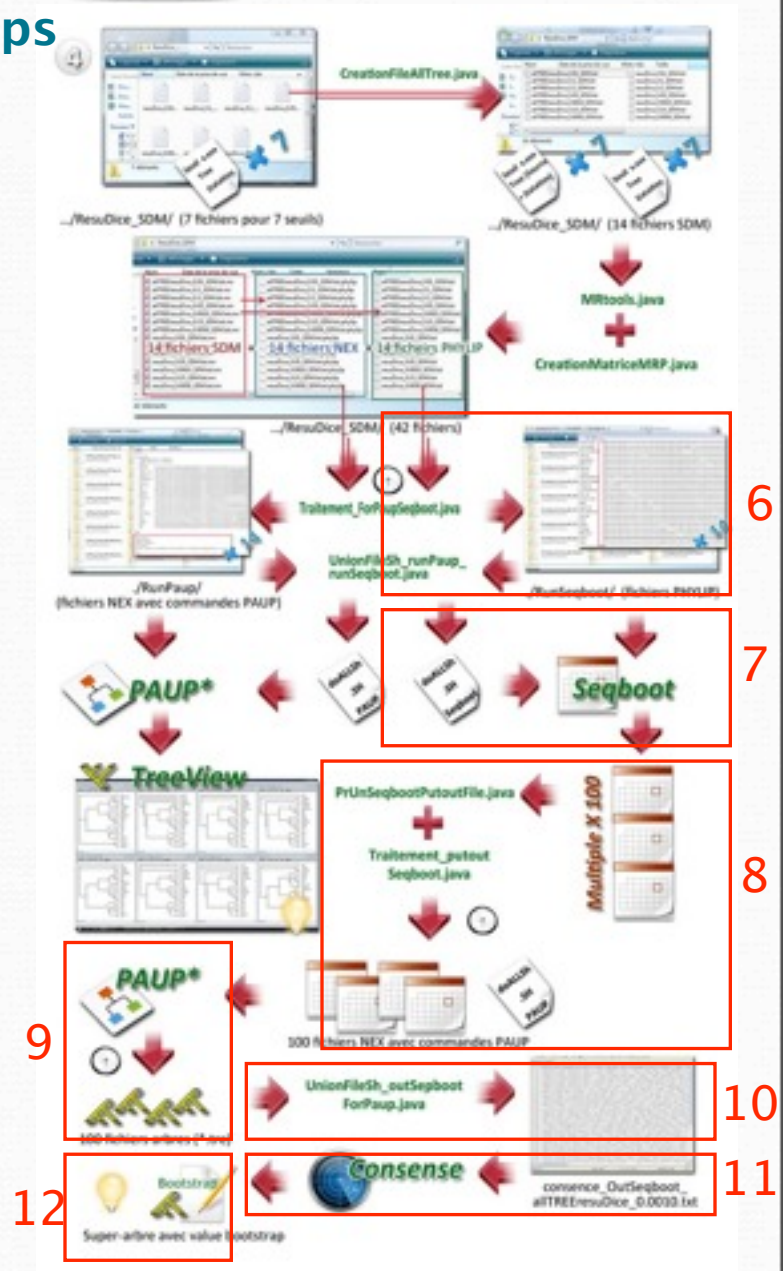
4. Traiter tous les fichiers par PAUP* pour construire les super-arbres par MRP

5. Consulter les arbres par TREEVIEW



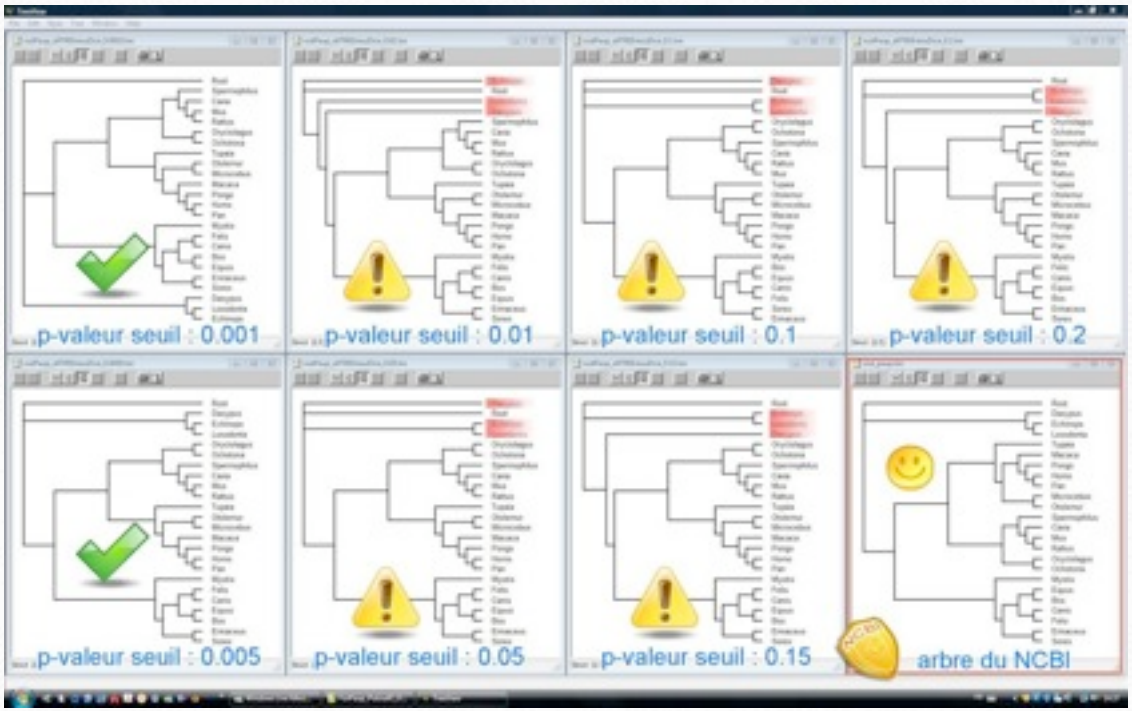
## ➢ Obtention des valeurs bootstraps

6. Pour chaque fichier d'output MRtools (Format PHYLIP), garder uniquement les 10 premières lettres des noms d'espèces et crée un fichier d'exécution runSeqboot.sh
7. Traiter tous les fichiers par Seqboot (PHYLIP)
   - Permettre de générer un jeu de données multiple (ici 100 fois multiple)
8. Enregistrer les 100 matrice 0-1 dans chaque fichier séparer, ajouter les commandes PAUP et crée un fichier d'exécution
9. Traiter tous les fichiers par PAUP*
10. Union tous les arbres obtenus dans un même fichier pour construire le consensus issu de l'analyse de bootstrap



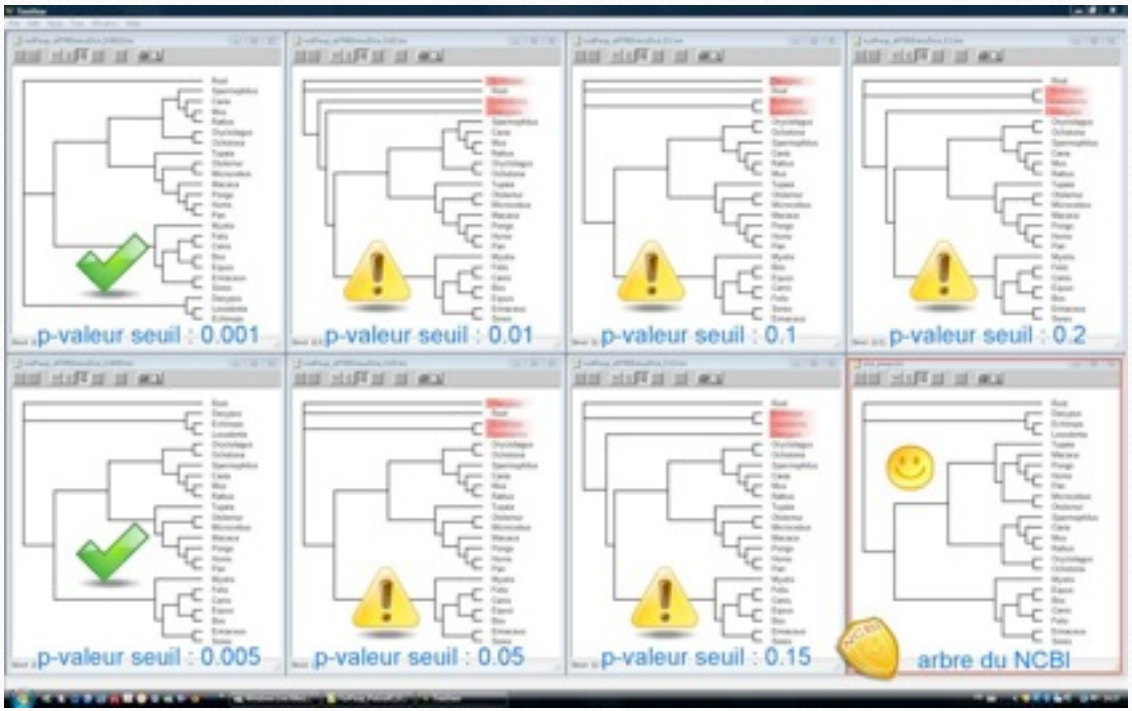
# III. Synthèse des résultats obtenus

## a. Analyser les super-arbres obtenus



# III. Synthèse des résultats obtenus

## a. Analyser les super-arbres obtenus




- Pour chaque méthode, comparer ces super-arbres (7 p-valeurs seuils) entre eux ainsi qu'à la taxonomie du NCBI.
- Augmenter la p-valeur seuil, certaines espèces (en rouge) sont mal positionnées(= augmenter le risque d'erreurs )
- La meilleure méthode peut diminuer cet effet.

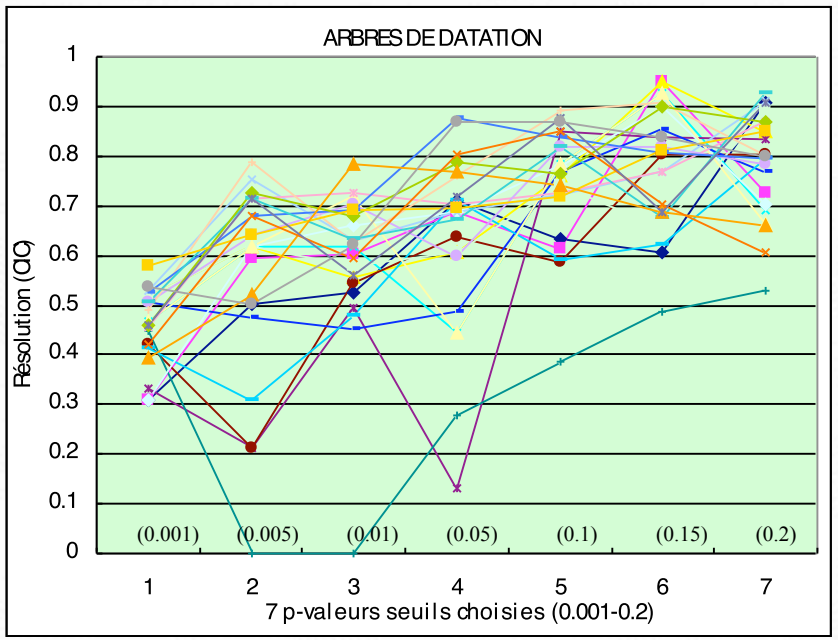
| 08/2009 | | méthode clique | p-valeur seuil | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 | 0.15 | 0.2 |
| 1er jeu de données | | V1 | ok | ok | ok | ok | ok | ok | ok |
| | | V2 | ok | ok | ok | ok | ok | ok | ok |
| | | V3 | ok | ok | ok | ok | ok | ok | ok |
| | | V4 | ok | ok | ok | ok | ok | ok | ok |
| | méthode R8s | méth. cliq. | arbres source + arbres de datation | | | | | | |
| deuxième jeu de données | LF - POWELL | V1 | ok | ok | X | X | X | X | X |
| | | V2 | ok | ok | X | ok | X | X | X |
| | | V3 | ok | ok | X | X | X | X | X |
| | | V4 | ok | ok | ok | X | X | ok | ok |
| | LF - TN | V1 | ok | ok | X | X | X | X | X |
| | | V2 | ok | ok | X | ok | X | X | X |
| | | V3 | ok | ok | X | X | X | X | X |
| | | V4 | ok | ok | ok | X | X | ok | ok |
| | NPRS - POWELL | V1 | ok | ok | ok | ok | X | ok | X |
| | | V2 | ok | ok | ok | X | ok | ok | X |
| | | V3 | ok | ok | ok | ok | ok | X | X |
| | | V4 | ok | ok | ok | ok | ok | ok | ok |
| | LF - QNEWT | V1 | ok | ok | ok | ok | X | X | X |
| | | V2 | ok | ok | ok | ok | X | X | X |
| | | V3 | ok | ok | ok | ok | X | X | X |
| | | V4 | ok | ok | ok | ok | ok | ok | ok |
| | PL - TN | V1 | ok | ok | ok | ok | X | X | ok |
| | | V2 | ok | ok | ok | ok | X | ok | ok |
| | | V3 | ok | ok | ok | ok | X | ok | X |
| | | V4 | ok | ok | ok | ok | X | ok | ok |
| | PL - QNEWT | V1 | ok | ok | ok | ok | ok | X | X |
| | | V2 | ok | ok | ok | ok | ok | X | X |
| | | V3 | ok | ok | ok | ok | ok | X | ok |
| | | V4 | ok | ok | ok | ok | ok | ok | ok |

- La meilleure méthode pour décomposer un graphe en cliques est la V4. Ici on considère que les pondérations sont des probabilités. Ceci correspond effectivement au fait que nous utilisons des p-valeurs pour valuer les arêtes des graphes construis.
- En ce qui concerne l'inférence des dates (méthode R8s), les méthodes qui aboutissent aux résultats les plus fiables sont PL/QNEWT et NPRS/POWELL.

# b. Analyser les résultats de CIC

- CIC : décrit la quantité de résolution d'un arbre. Pour chaque arbre la valeur obtenue est comprise entre 0 et 1, une valeur proche de 1 indique une meilleure résolution
- Analyser les arbres produits par la méthode exploitant les datations
  - suivant les valeurs des paramètres, les datations apportent un nombre différent d'informations supplémentaires pour la construction d'un super-arbre.
  - lorsque l'on augmente la p-valeur seuil, les arbres sont mieux

| 08/2009 | | méthode | arbres de datation | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | p-valeur seuil | clique | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 | 0.15 | 0.2 |
| 1er jeu de données | | V1 | 0.81176 | 0.81176 | 0.81176 | 0.81176 | 0.81176 | 0.81176 | 0.81176 |
| | | V2 | 0.81176 | 0.81176 | 0.81176 | 0.81176 | 0.81176 | 0.81176 | 0.81176 |
| | | V3 | 0.81176 | 0.81176 | 0.81176 | 0.81176 | 0.81176 | 0.81176 | 0.81176 |
| | | V4 | 0.81176 | 0.81176 | 0.81176 | 0.81176 | 0.81176 | 0.81176 | 0.81176 |
| | méthode R8s | méth. cliq. | arbres de datation | | | | | | |
| deuxième jeu de données | LF - POWELL | V1 | 0.30873 | 0.50206 | 0.52667 | 0.71077 | 0.63142 | 0.60564 | 0.9082 |
| | | V2 | 0.30873 | 0.59363 | 0.60144 | 0.68616 | 0.61462 | 0.94961 | 0.72757 |
| | | V3 | 0.51886 | 0.61824 | 0.55489 | 0.6072 | 0.76205 | 0.94961 | 0.85 |
| | | V4 | 0.4654 | 0.61824 | 0.61824 | 0.44346 | 0.78358 | 0.925 | 0.69001 |
| | LF - QNEWT | V1 | 0.33333 | 0.2109 | 0.49553 | 0.13108 | 0.85 | 0.83705 | 0.8332 |
| | | V2 | 0.4217 | 0.2109 | 0.54347 | 0.63695 | 0.58525 | 0.80345 | 0.80345 |
| | | V3 | 0.44631 | 7.97E-18 | 7.97E-18 | 0.27987 | 0.38584 | 0.48771 | 0.53051 |
| | | V4 | 0.5068 | 0.4754 | 0.45079 | 0.48771 | 0.76679 | 0.85384 | 0.76679 |
| | LF - TN | V1 | 0.41271 | 0.30873 | 0.47873 | 0.71077 | 0.59002 | 0.62243 | 0.79564 |
| | | V2 | 0.30873 | 0.61824 | 0.66156 | 0.68616 | 0.77884 | 0.89922 | 0.70296 |
| | | V3 | 0.51886 | 0.61824 | 0.57834 | 0.6072 | 0.74218 | 0.65603 | 0.89141 |
| | | V4 | 0.4654 | 0.61824 | 0.71976 | 0.44346 | 0.78358 | 0.925 | 0.6654 |
| | NPRS - POWELL | V1 | 0.52794 | 0.75383 | 0.62989 | 0.68616 | 0.72024 | 0.76679 | 0.91205 |
| | | V2 | 0.45412 | 0.71243 | 0.72757 | 0.70296 | 0.72757 | 0.76679 | 0.87064 |
| | | V3 | 0.50527 | 0.64217 | 0.70296 | 0.59821 | 0.82025 | 0.82025 | 0.78358 |
| | | V4 | 0.49001 | 0.78577 | 0.62628 | 0.76116 | 0.89141 | 0.9082 | 0.80345 |
| | PL - QNEWT | V1 | 0.52398 | 0.67883 | 0.69001 | 0.87578 | 0.83705 | 0.80729 | 0.79564 |
| | | V2 | 0.50718 | 0.71243 | 0.6318 | 0.67102 | 0.82025 | 0.67644 | 0.925 |
| | | V3 | 0.46025 | 0.72538 | 0.67835 | 0.78783 | 0.76589 | 0.90039 | 0.8668 |
| | | V4 | 0.57834 | 0.63923 | 0.69001 | 0.69473 | 0.71757 | 0.81244 | 0.85 |
| | PL - TN | V1 | 0.39365 | 0.52013 | 0.78358 | 0.76898 | 0.74218 | 0.68616 | 0.65964 |
| | | V2 | 0.42052 | 0.67883 | 0.5963 | 0.80345 | 0.85 | 0.70296 | 0.60528 |
| | | V3 | 0.46025 | 0.71243 | 0.55963 | 0.71976 | 0.87461 | 0.68616 | 0.9082 |
| | | V4 | 0.53502 | 0.50166 | 0.61975 | 0.8668 | 0.8668 | 0.83705 | 0.79948 |



- Pour ce jeu, les arbres sources initiaux permettent déjà d'inférer un arbre binaire, en accord avec la taxonomie du NCBI.

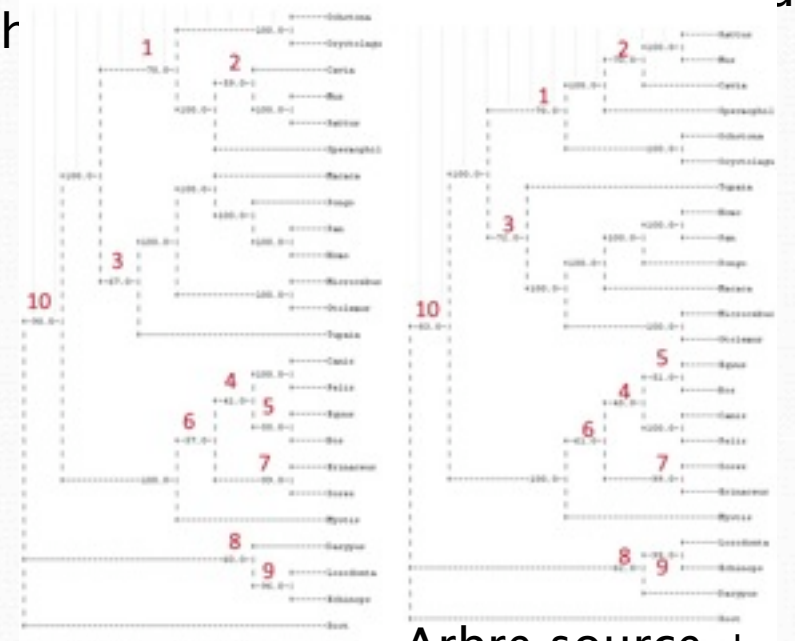
| 08/2009 | | méthode | arbres source + arbres de datation | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | p-valeur seuil | clique | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 | 0.15 | 0.2 |
| 1er jeu de données | | V1 | 0.95294 | 0.95294 | 0.95294 | 0.95294 | 0.95294 | 0.95294 | 0.95294 |
| | | V2 | 0.95294 | 0.95294 | 0.95294 | 0.95294 | 0.95294 | 0.95294 | 0.95294 |
| | | V3 | 0.95294 | 0.95294 | 0.95294 | 0.95294 | 0.95294 | 0.95294 | 0.95294 |
| | | V4 | 0.95294 | 0.95294 | 0.95294 | 0.95294 | 0.95294 | 0.95294 | 0.95294 |
| | méthode R8s | méth. cliq. | arbres source + arbres de datation | | | | | | |
| deuxième jeu de données | LF - POWELL | V1 | 0.9832 | 0.9832 | 0.9832 | 0.9586 | 0.9586 | 0.9832 | 0.9832 |
| | | V2 | 0.9832 | 0.9832 | 0.92884 | 0.9832 | 0.9586 | 0.9832 | 0.9586 |
| | | V3 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 |
| | | V4 | 0.9832 | 0.9832 | 0.9832 | 0.9586 | 0.9586 | 0.9832 | 0.9832 |
| | LF - QNEWT | V1 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9586 | 0.9832 |
| | | V2 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9586 | 0.9832 |
| | | V3 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9586 | 0.9586 | 0.9832 |
| | | V4 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 |
| | LF - TN | V1 | 0.9832 | 0.9832 | 0.9832 | 0.9586 | 0.9586 | 0.9832 | 0.9832 |
| | | V2 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9586 | 0.9832 | 0.9832 |
| | | V3 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 |
| | | V4 | 0.9832 | 0.9832 | 0.9832 | 0.9586 | 0.9586 | 0.9832 | 0.9832 |
| | NPRS - POWELL | V1 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9586 | 0.9832 | 0.9586 |
| | | V2 | 0.9832 | 0.9832 | 0.9832 | 0.9586 | 0.9832 | 0.9832 | 0.9586 |
| | | V3 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9586 | 0.9832 |
| | | V4 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 |
| | PL - QNEWT | V1 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9586 | 0.9586 |
| | | V2 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9586 | 0.9586 |
| | | V3 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9586 | 0.9832 |
| | | V4 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.91205 | 0.94961 |
| | PL - TN | V1 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.92884 | 0.92884 | 0.9832 |
| | | V2 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.92884 | 0.9832 | 0.9832 |
| | | V3 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.92884 | 0.9832 | 0.9586 |
| | | V4 | 0.9832 | 0.9832 | 0.9832 | 0.9832 | 0.9586 | 0.9832 | 0.9832 |

| arbres source | |
|---|---|
| premier jeu de données | deuxième jeu de données |
| 0.952939 | 0.98323 |

- Nous avons pu vérifier qu'ajouter les arbres de datations aux arbres sources de départ pour inférer un super-arbre ne change pas (ou très peu et rarement) le fait qu'il soit complètement résolus.

# c. Analyser les résultats de Bootstrap

- Dans cet exemple, les valeurs bootstrap augmentent, mais pas beaucoup.
- Les tests de bootstrap sont  fini, nous allons créer un programme pour traiter ces quelques milliers de valeurs.
- Pour conclure, il faut aussi tester les autres résultats. Ils seront disponibles dans quelques jours.
- Il reste aussi à analyser un jeu de données avec moins de ch



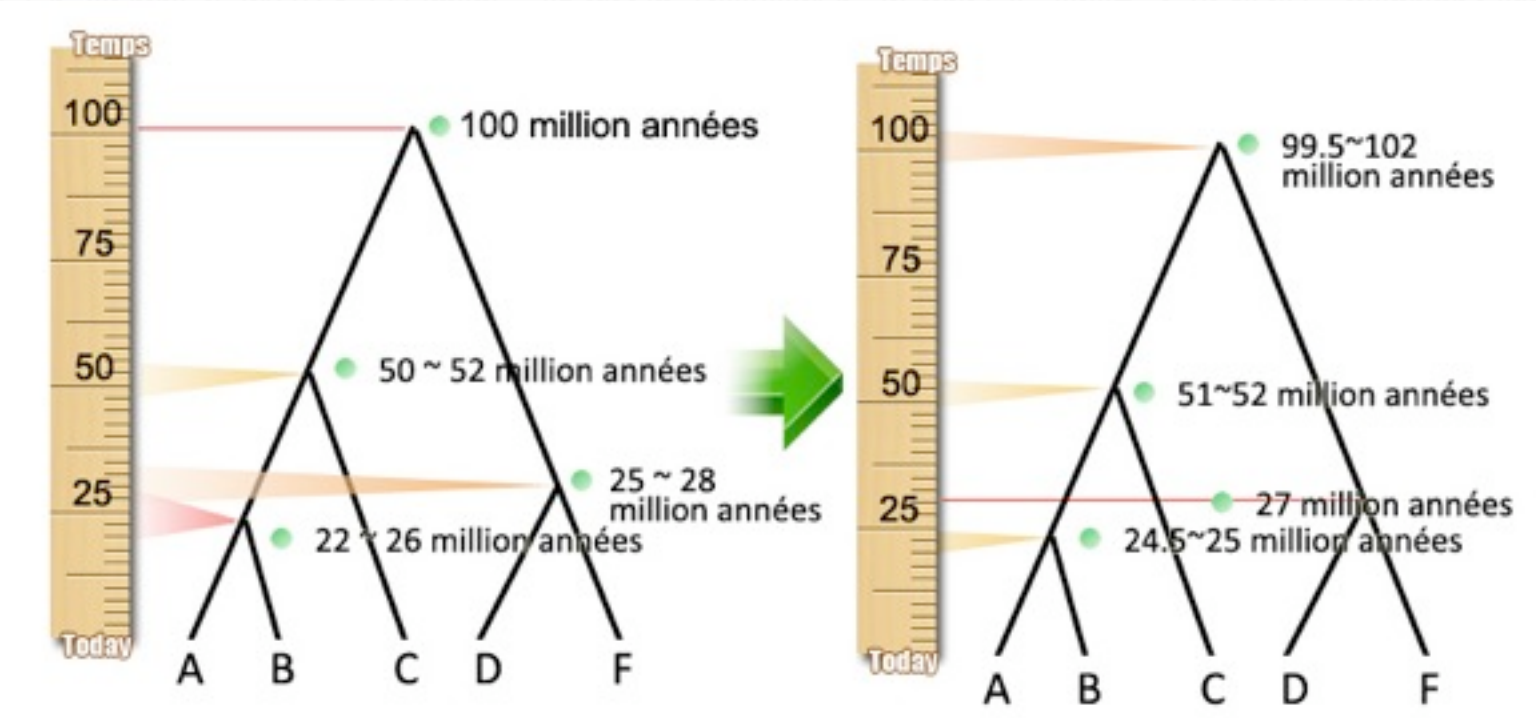
Arbres source

Arbre source + arbres de datations

| Pl / Qnewt / v4 / 0.001 | | |
|---|---|---|
| nœud | source | nouv. méth. |
| 1 | 70 | 70 |
| 2 | 59 | 70 |
| 3 | 67 | 72 |
| 4 | 41 | 45 |
| 5 | 50 | 51 |
| 6 | 57 | 61 |
| 7 | 99 | 99 |
| 8 | 60 | 62 |
| 9 | 96 | 99 |
| 10 | 90 | 83 |

Méthode R8s : Pl/Qnewt
Clique : V4
P-valeur seuil : 0.001

# Amélioration possible




- Changer la façon de fixer la date du nœud, au lieu de fixer dans la racine, fixer dans un couple de espèces quelconque ou un couple connu plus vers milieu, ça va augmenter la précision pour prédire les âge des nœuds autour de lui.

# Conclusion

- Les résultats obtenus dans ce stage de recherche vont faire l'objet d'une publication scientifique dans les prochaines semaines.

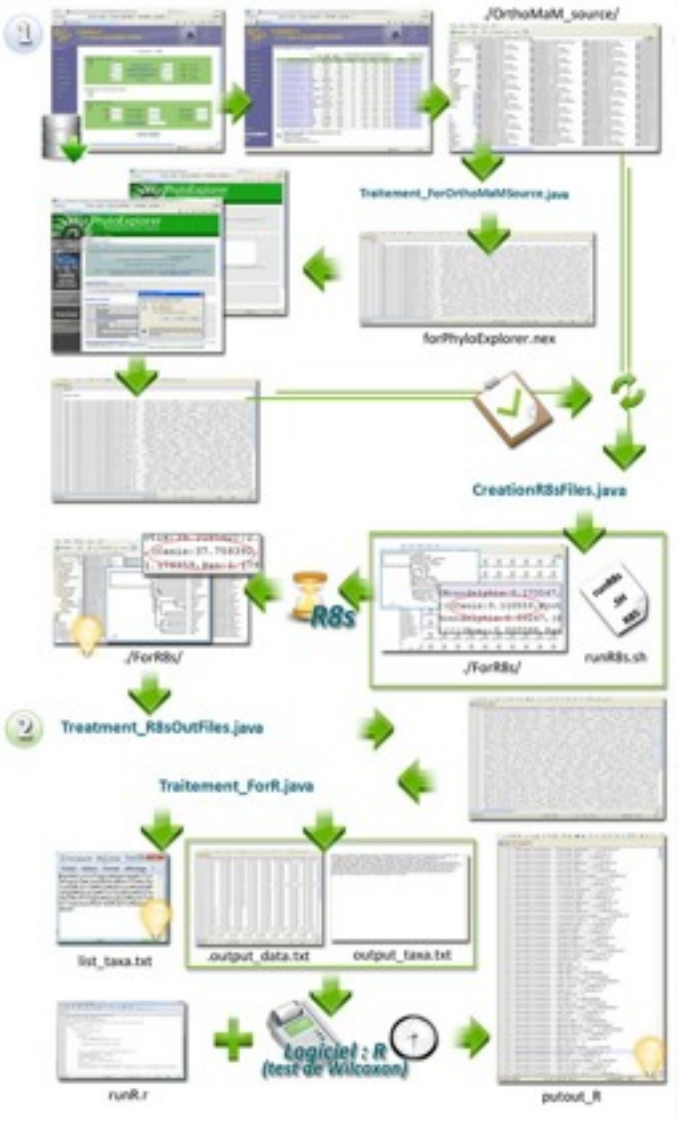
- Lors de mon stage de deuxième année, j'ai découvert le monde de la recherche et ce stage, m'a permis d'
  - approfondir mes connaissances dans le domaine phylogénétique,
  - enrichir mon expérience de développement en informatique,
  - discuter et résoudre des problèmes de recherche,
  - utiliser de nombreuses technologies que j'ai apprises lors de la formation de Master.
- La réalisation de ce stage reste donc pour moi une très bonne expérience dans le domaine de la recherche et un

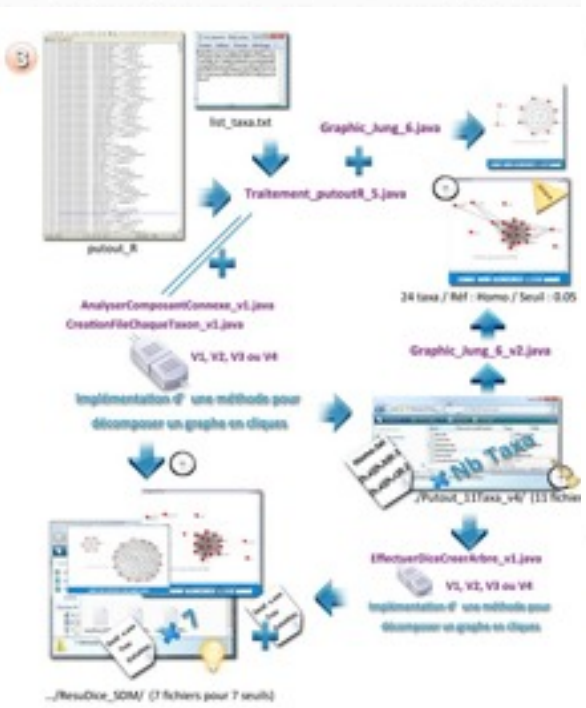
# Référence

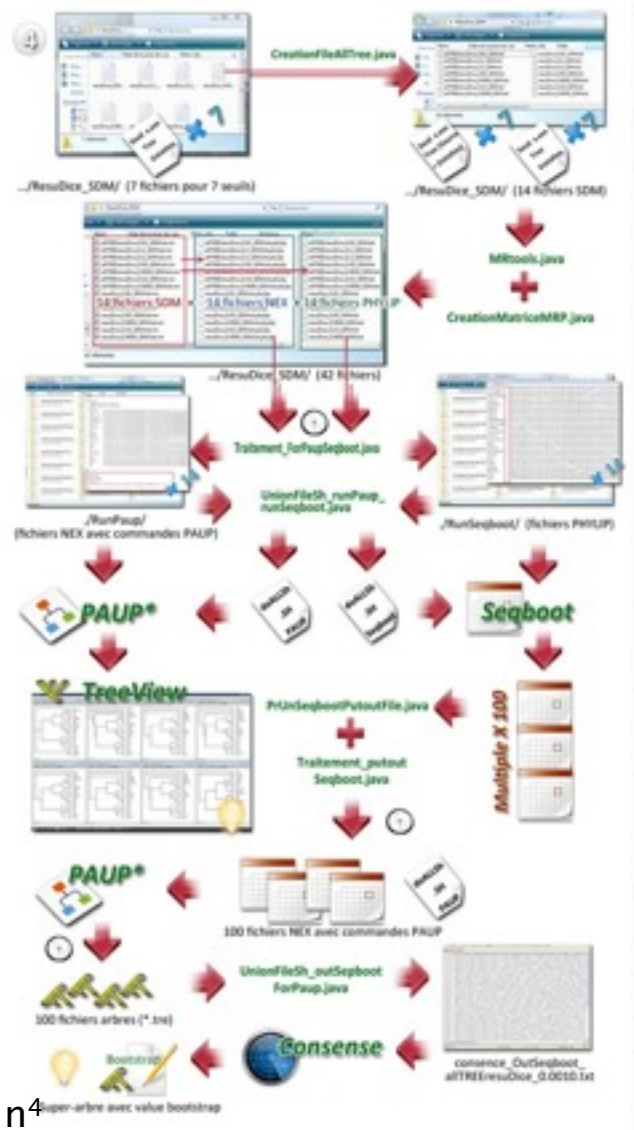

- Adams, 1972. Consensus techniques and the comparison of taxonomic trees. Systematic Zoology.
- Angelelli et al. 2008, Two local dissimilarity measures for weighted graphs with application to protein interaction networks. Advances in Data Analysis and Classification, 2:3-16.
- Baum, 1992. Combining trees as a way of combining data sets for phylogenetic inference. Taxon.
- Berry et al., 2004 Maximum Agreement and Compatible Supertrees Lecture notes in computer science ISSN 0302-9743
- Bininda-Emonds et al., 2002. The (super)tree of life: procedures, problems, and prospects. Annu. Rev. Ecol. Syst.
- Bininda-Emonds, 1998. Properties of Matrix Representation with Parcimony Analyses. Systematic Biology.
- Bininda-Emonds, 2000. MRP supertree construction in the consensus setting. DIMACS Séries in Discrete Mathematics and Theoretical Computer Science.
- Bininda-Emonds, 2001. Assessment of the accuracy of Matrix Representation with Parcimony Analysis Supertree Sonstruction. Systematic Biology.
- Brun C et al., 2003 Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biol, 5, R6.
- Chen et al., 2002. Flipping: a supertree construction method. DIMACS Series in Discrete Mathematics and Theoretical Computer Sciences.
- Chen et al., 2003. Supertrees by flipping. COCOON.
- Criscoul A. , 2006 MRtools site : http://www.lirmm.fr/~criscuol/
- Criscuol A.et al., 2006. SDM: a fast distance-based approach for (super) tree building in phylogenomics. Systematic Biology 55 (5): 740-755.
- Dice 1945 Measures of the amount of ecological association between species. Ecology 26:297-302
- Eulenstein et al., 2004. Performance of flip supertrees with a heuristic algorithm. Systematic Biology.
- Gordon, A., 1986 Consensus supertrees: the synthesis of rooted trees containing overlapping sets of leaves. J. Classification 3 (1986), pp. 335-348.
- Guénoche, 2005 Comparison of algorithms in graph partitioning, ALIO/EURO conference on Combinatorial Optimization, Paris, submitted.
- Ragan, 1992. Phylogenetic inference based on matrix representation of trees. Mol. Phyl. Evol.
- Ranwez V. et al., 2007 OrthoMaM : A database of orthologous genomic markers for placental mammal phylogenetics. BMC Evolutionary Biology, 7 : 241
- Ranwez V. et al., 2009 PhyloExplorer: a web server to validate, explore and query phylogenetic trees. BMC Evolutionary Biology 9:108

# Merci de votre Attention!

# Discussion




Ajouter un Taxon : Matrice1 $l*(n+1)^2$

Matrice2 (n

Ajouter un Arbre : Matrice1 (l +1)*$n^2$

A+1)$^4$

Martice2 $n^4$

- Par conséquent, on possède beaucoup de génomes pour certains groupes, comme par exemple les Mammifères, tandis que d'autres groupes sont peu séquencés (comme par exemple les oiseaux, les amphibiens, ou encore les tuniciers.

# Différentes combinaisons d'algorithmes disponibles

**Table 1**. Algorithms implemented for various methods.

| Method | Constraints | Non-extant terminals | POWELL | TN | QNEWT |
|---|---|---|---|---|---|
| | | | | | |
| LF | no | no | yes | yes | yes |
| | no | yes | yes | yes | no |
| | yes | no | yes | yes | no |
| | yes | yes | yes | yes | no |
| | | | | | |
| LF (local) | no | no | yes | no | no |
| | no | yes | yes | no | no |
| | yes | no | yes | no | no |
| | yes | yes | yes | no | no |
| | | | | | |
| NPRS | no | no | yes | no | no |
| | no | yes | yes | no | no |
| | yes | no | yes | no | no |
| | yes | yes | yes | no | no |
| | | | | | |
| PL | no | no | yes | yes | yes |
| | no | yes | yes | yes | no |
| | yes | no | yes* | yes | no |
| | yes | yes | yes* | yes | no |

* but not cross-validation!