Robustesse des arbres phylogénétiques: échantillonage de taxon

M. Mariadassou A. Bar-Hen H. Kishino

Laboratoire MAP5
Université Paris Descartes

Août 2009
Rencontres des Jeunes Statisticiens

Phylogénie Moléculaire

But de la phylogénie moléculaire :

Utiliser des macromolécules biologiques (ADN, ARN, protéines) pour étudier l'Évolution.

Histoire évolutive :

- Accumulation de différences dans des sous-populations ;
- Spéciation ou création de nouvelles espèces ;
- L'évolution a une forme d'arbre.

Philosophie sous-jacente :

- Espèces très proches : molécules très similaires ;
- Espèces plus distantes : molécules moins similaires,
- Utiliser la ressemblance pour reconstruire l'arbre d'évolution.

Phylogénie Moléculaire

But de la phylogénie moléculaire :

Utiliser des macromolécules biologiques (ADN, ARN, protéines) pour étudier l'Évolution.

Histoire évolutive :

- Accumulation de différences dans des sous-populations ;
- Spéciation ou création de nouvelles espèces ;
- L'évolution a une forme d'arbre.

Philosophie sous-jacente :

- Espèces très proches : molécules très similaires ;
- Espèces plus distantes : molécules moins similaires,
- Utiliser la ressemblance pour reconstruire l'arbre d'évolution.

Phylogénie Moléculaire

But de la phylogénie moléculaire :

Utiliser des macromolécules biologiques (ADN, ARN, protéines) pour étudier l'Évolution.

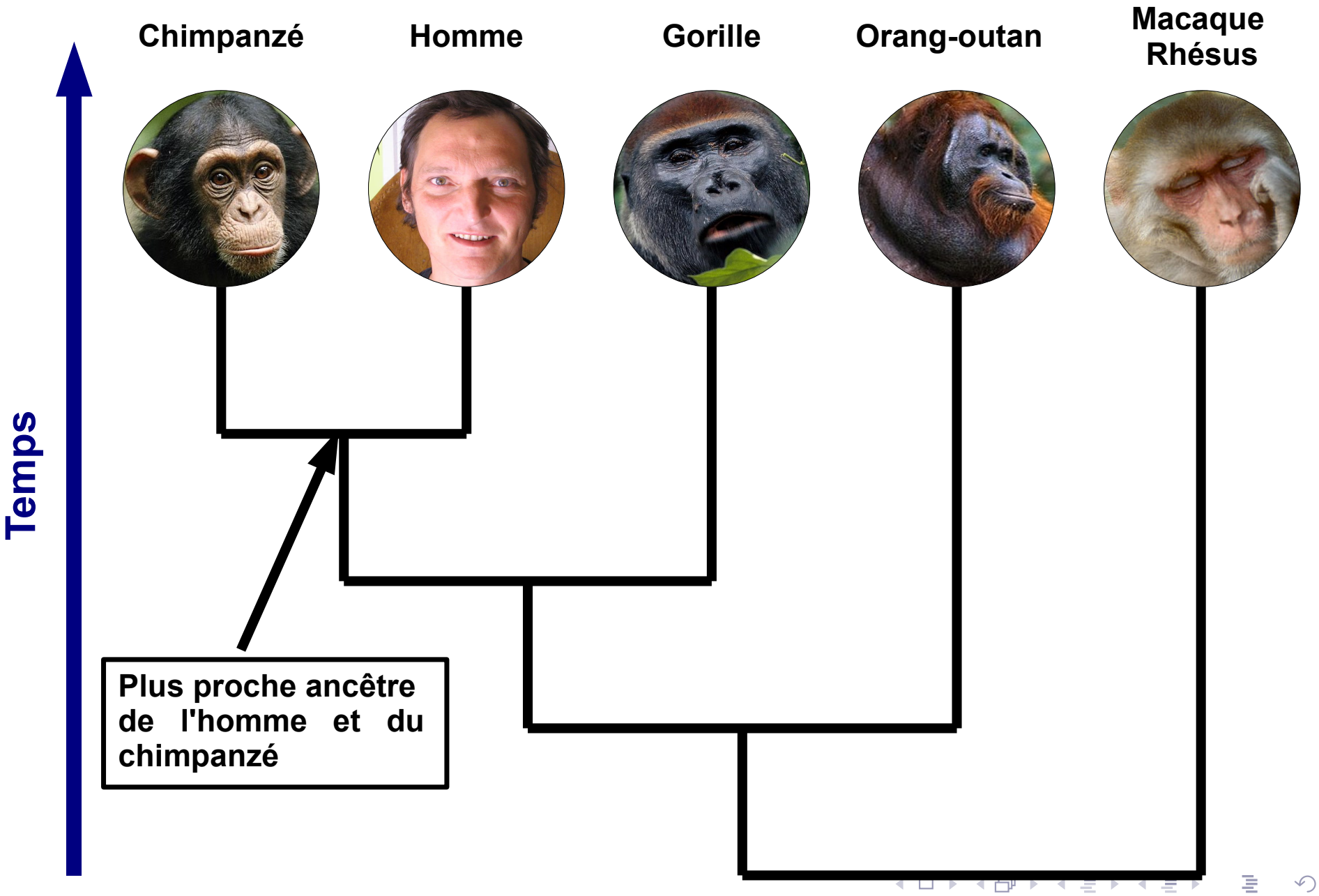
Histoire évolutive :

- Accumulation de différences dans des sous-populations ;
- Spéciation ou création de nouvelles espèces ;
- L'évolution a une forme d'arbre.

Philosophie sous-jacente :

- Espèces très proches : molécules très similaires ;
- Espèces plus distantes : molécules moins similaires,
- Utiliser la ressemblance pour reconstruire l'arbre d'évolution.

Exemple de phylogénie : grands singes



La phylogénie en pratique

Objectif :

- Saint-Graal : reconstruire l'Arbre de la Vie ;
- En pratique : phylogénie d'un groupe plus limité d'espèces ;

Applications :

- Bioinformatique (annotations de gènes, biologie comparative, . . .) ;
- Protection de l'environnement (mesure de biodiversité) ;
- Épidémiologie (SRAS, HIV, H1N1)

La phylogénie en pratique

Objectif :

- Saint-Graal : reconstruire l'Arbre de la Vie ;
- En pratique : phylogénie d'un groupe plus limité d'espèces ;

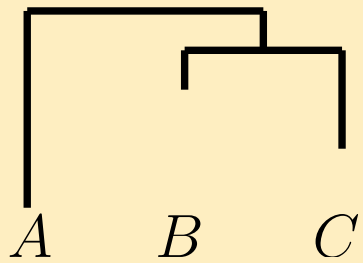
Applications :

- Bioinformatique (annotations de gènes, biologie comparative, . . .) ;
- Protection de l'environnement (mesure de biodiversité) ;
- Épidémiologie (SRAS, HIV, H1N1)

Reconstruction et problèmes

Deux niveaux de reconstruction :

- Reconstruire la phylogénie :
  - Topologie,
  - Longueurs de branches.



- Reconstruire les états ancestraux.

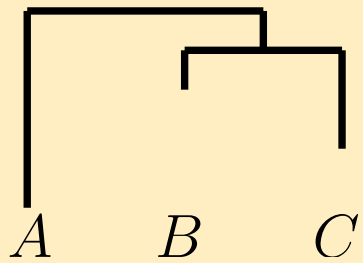
Difficultés :

- Évolution : événement unique ;
- Information génétique disponible uniquement sur les espèces encore existantes ;
- Peu d'observation directes sur le processus évolutif.

Reconstruction et problèmes

Deux niveaux de reconstruction :

- Reconstruire la phylogénie :
  - Topologie,
  - Longueurs de branches.



- Reconstruire les états ancestraux.

Difficultés :

- Évolution : événement unique ;
- Information génétique disponible uniquement sur les espèces encore existantes ;
- Peu d'observation directes sur le processus évolutif.

Les données

Collection : Choisir un gène/protéine présent chez toutes les espèces, le séquencer et aligner les séquences.

Exemple :

- Alignement $\mathcal{X} = (X_{ij})$ de taille $s \times n$ (6 espèces $\times$ 10 sites)

Baleine fin.	M	N	E	**N**	L	F	A	P	F
Baleine bleue	M	N	E	**N**	L	F	A	P	F
Chimpanzé	M	N	E	**N**	L	F	A	S	F
Bonobo	M	N	E	**N**	L	F	A	S	F
Gorille	M	N	E	**N**	L	F	A	S	F
Orang-outan	M	N	E	**D**	L	F	T	P	F

- $\mathcal{X}_{24} =$ **N**,
- 4ème site : $\mathbf{X}_4 =$ (**NNNNND**)$'$,
- 2$^{\text{nd}}$ espèce (Baleine bleue) : $\mathbf{X}^{(2)} = MNENLFAPFM$.

Méthodes d'estimation

Trois familles de méthodes :

- Choisir un critère : parcimonie (MP), vraisemblance (IB,MV), moindres carrés (NJ) ;
- Estimer l'arbre qui optimise ce critère.

Méthodes d'estimation

Trois familles de méthodes :

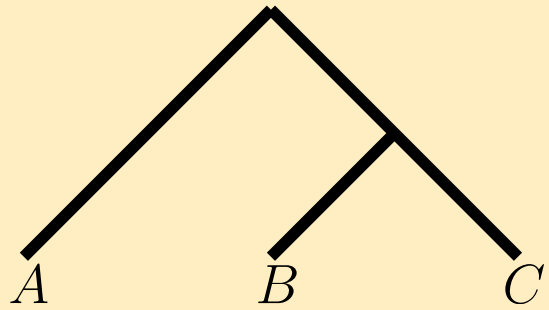
- Choisir un critère : parcimonie (MP), vraisemblance (IB,MV), moindres carrés (NJ) ;
- Estimer l'arbre qui optimise ce critère.

Principe :

Alignement $\longrightarrow$ **Phylogénie**

A	C	C	T	T
B	G	G	A	A
C	G	G	A	C

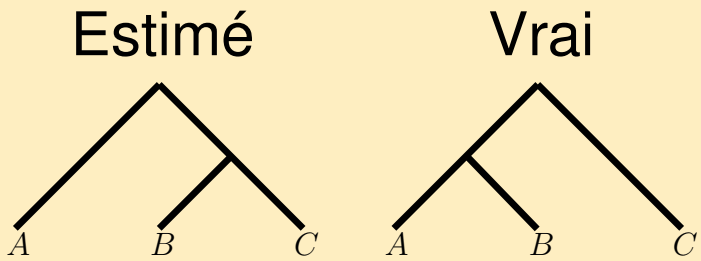
MP, NJ
MV, IB
...



Fin de l'histoire ?

Validation de la méthode :

- Comparer l'arbre estimé au vrai arbre pour mesurer la qualité de l'estimation,



- Mais le vrai arbre est inconnu !

Problème de validation :

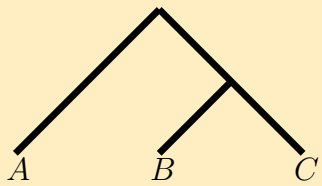
- Quel crédit accorder à l'arbre estimé ?
- Quelles portions de l'arbre sont dignes de confiance ?

Fin de l'histoire ?

Validation de la méthode :

- Comparer l'arbre estimé au vrai arbre pour mesurer la qualité de l'estimation,

Estimé



Vrai

?

- Mais le vrai arbre est inconnu !

Problème de validation :

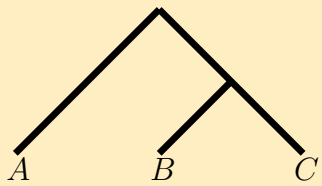
- Quel crédit accorder à l'arbre estimé ?
- Quelles portions de l'arbre sont dignes de confiance ?

Fin de l'histoire ?

Validation de la méthode :

- Comparer l'arbre estimé au vrai arbre pour mesurer la qualité de l'estimation,

Estimé



Vrai

?

- Mais le vrai arbre est inconnu !

Problème de validation :

- Quel crédit accorder à l'arbre estimé ?
- Quelles portions de l'arbre sont dignes de confiance ?

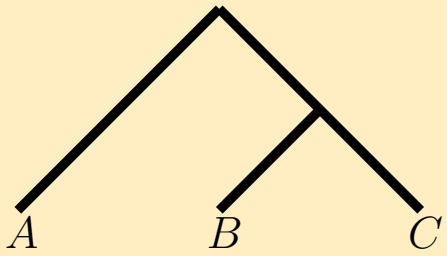
Valeurs de bootstrap : la théorie

Jeu de données d'origine :

Alignement

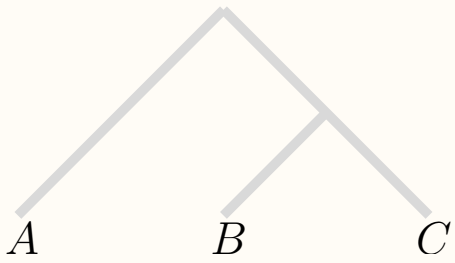
A	A	C	T	T
B	G	G	A	T
C	G	G	C	C

Phylogénie

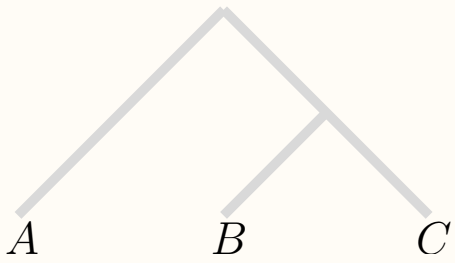


Jeux de données bootstrap :

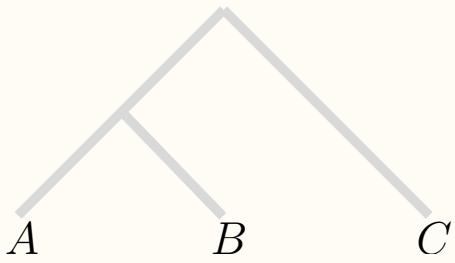
A	A	C	T	C
B	G	G	A	G
C	G	G	C	G



A	C	A	T	A
B	G	G	A	G
C	G	G	C	G



A	T	T	T	T
B	A	T	A	T
C	C	C	C	C



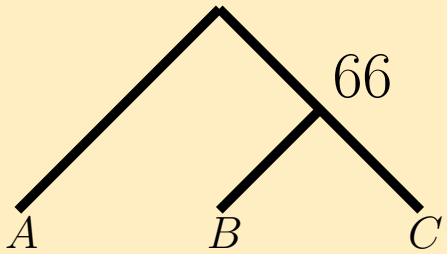
Valeurs de bootstrap : la théorie

Jeu de données d'origine :

Alignement

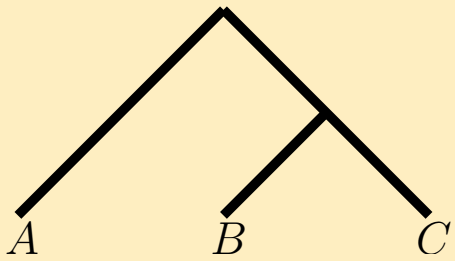
A	A	C	T	T
B	G	G	A	T
C	G	G	C	C

Phylogénie

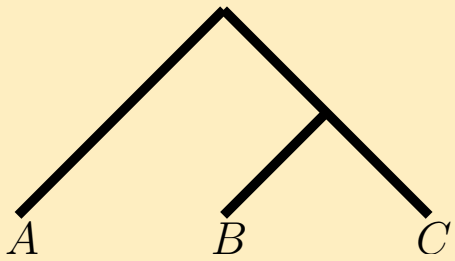


Jeux de données bootstrap :

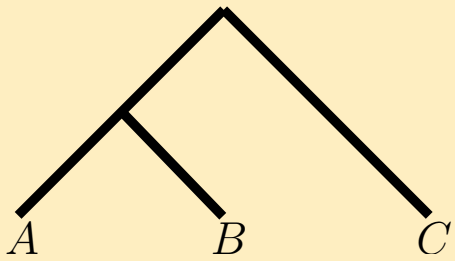
A	A	C	T	C
B	G	G	A	G
C	G	G	C	G



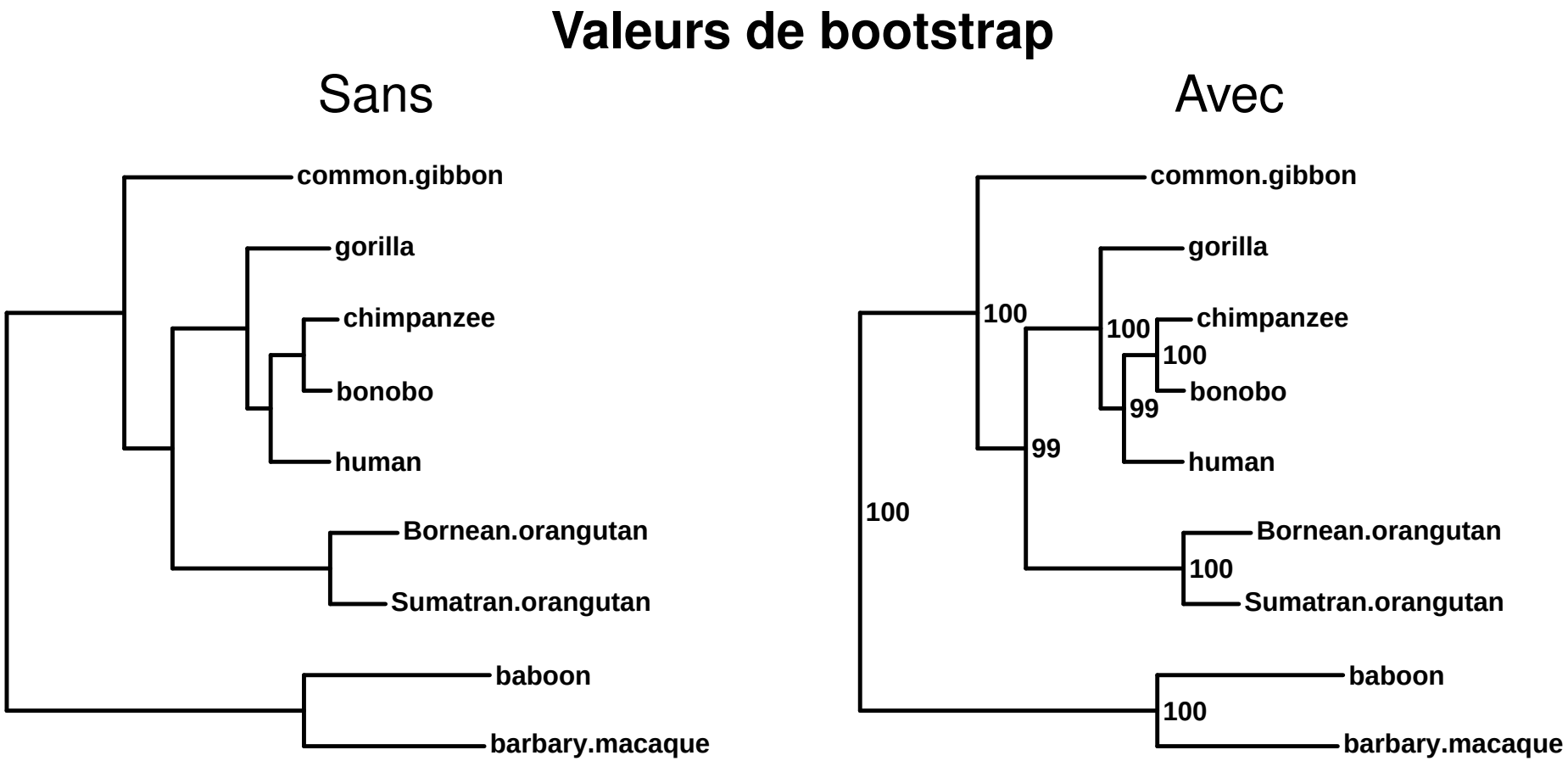
A	C	A	T	A
B	G	G	A	G
C	G	G	C	G



A	T	T	T	T
B	A	T	A	T
C	C	C	C	C



Valeurs de bootstrap : un exemple



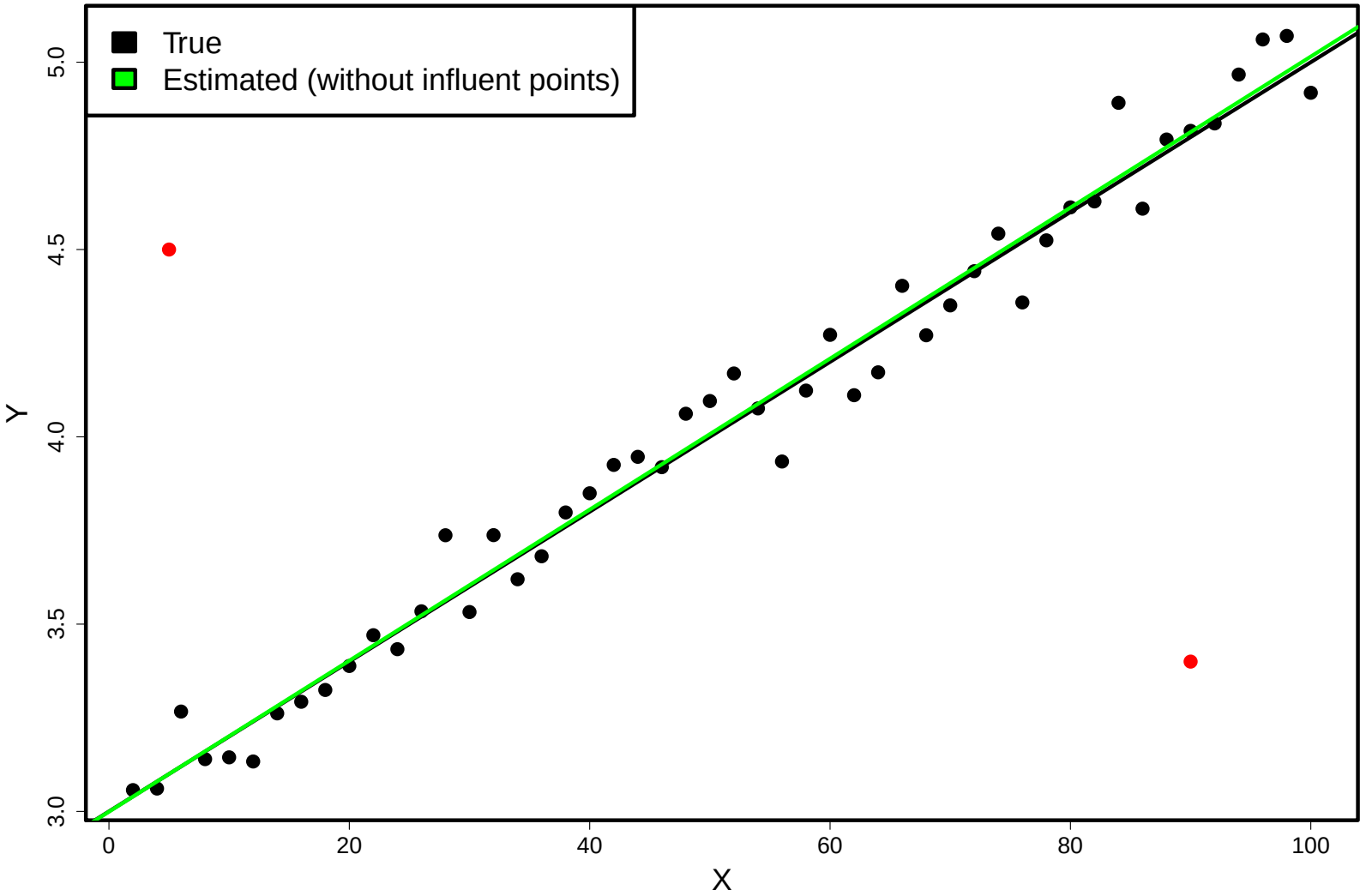
D'autres facteurs de variabilité

- Bootstrap mesure la variabilité induite par l'échantillonnage de sites ;
- Bootstrap aveugle à :
  - Erreurs d'alignement (sites aberrants) ;
  - Échantillonnage d'espèces (espèces voyous) ;
  - Erreur dans le modèle ;
- Sensibilité de l'arbre à ces facteurs ?
- Besoin d'autres mesures (locales) de variabilité.

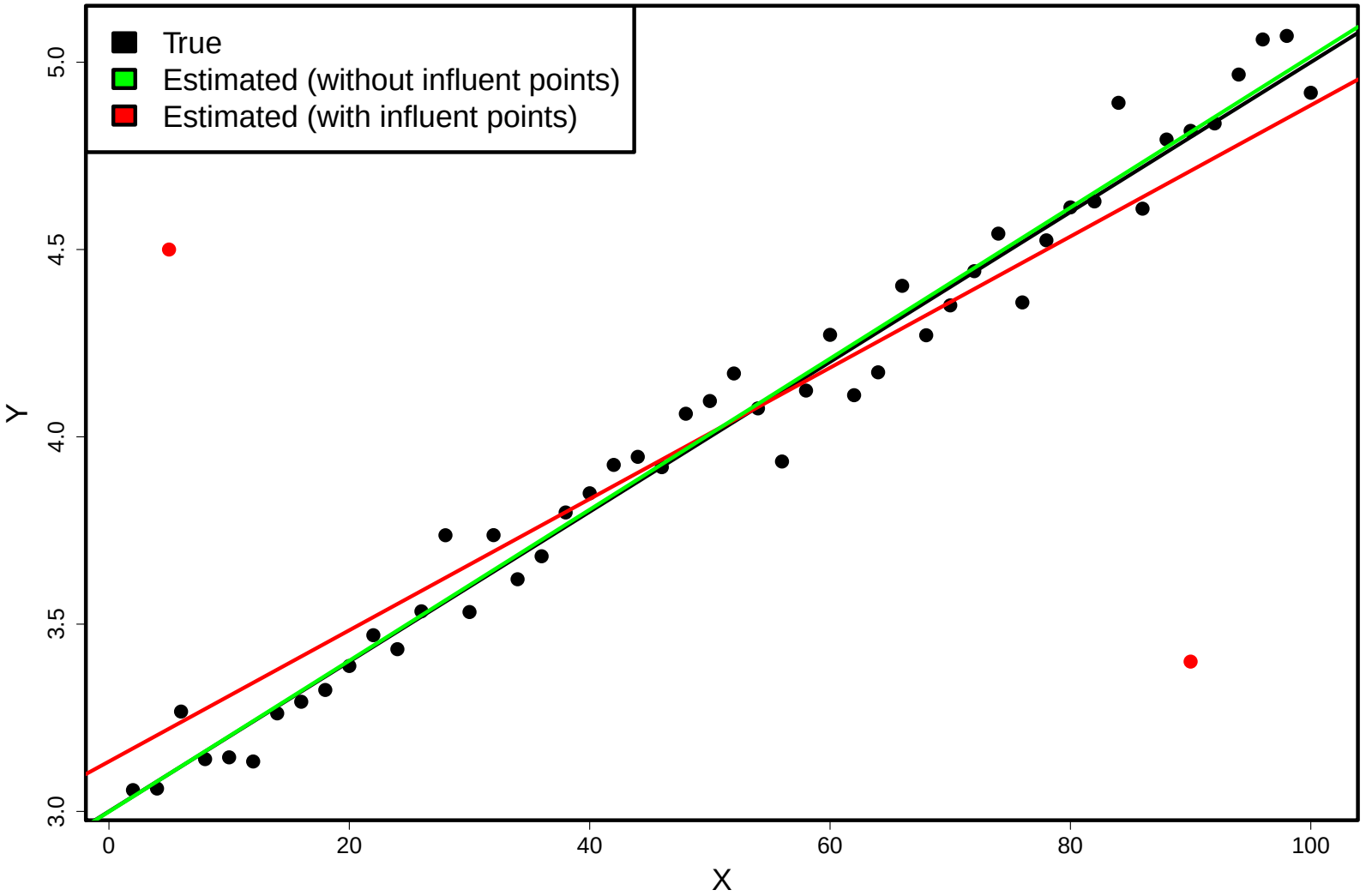
D'autres facteurs de variabilité

- Bootstrap mesure la variabilité induite par l'échantillonnage de sites ;
- Bootstrap aveugle à :
  - Erreurs d'alignement (sites aberrants) ;
  - Échantillonnage d'espèces (espèces voyous) ;
  - Erreur dans le modèle ;
- Sensibilité de l'arbre à ces facteurs ?
- Besoin d'autres mesures (locales) de variabilité.

Un exemple simple en régression linéaire



Un exemple simple en régression linéaire



Taxon Influence Index : objectif et heuristique

Taxon Influence Index (TII)

- **Objectif :** Étudier la stabilité de l'arbre vis-à-vis des espèces,
- **Justification :** Certaines espèces présentent des caractéristiques évolutives, non prises en compte par le modèle, qui peuvent biaiser la phylogénie estimée.
- **Méthode :**
  - Estimer la phylogénie T sur jeu complet d'espèces,
  - Retirer les espèces une par une et estimer le nouvel arbre T_i sur le nouveau jeu d'espèces,
  - Évaluer la différence entre T et T_i par $d(T, T_i)$.

Taxon Influence Index : objectif et heuristique

Taxon Influence Index (TII)

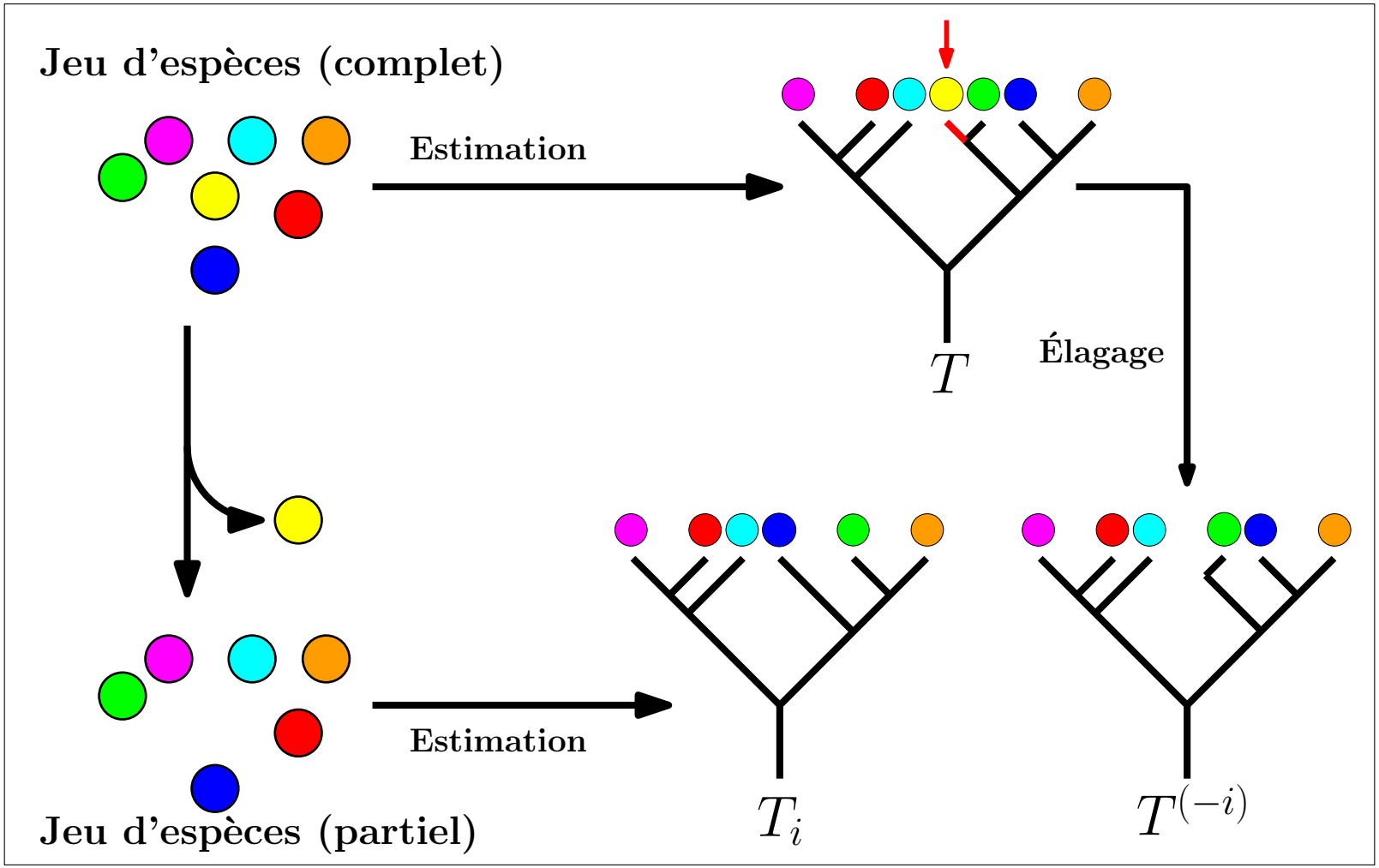
- **Objectif :** Étudier la stabilité de l'arbre vis-à-vis des espèces,
- **Justification :** Certaines espèces présentent des caractéristiques évolutives, non prises en compte par le modèle, qui peuvent biaiser la phylogénie estimée.
- **Méthode :**
  - Estimer la phylogénie T sur jeu complet d'espèces,
  - Retirer les espèces une par une et estimer le nouvel arbre T_i sur le nouveau jeu d'espèces,
  - Évaluer la différence entre T et T_i par $d(T, T_i)$.

Taxon Influence Index : objectif et heuristique

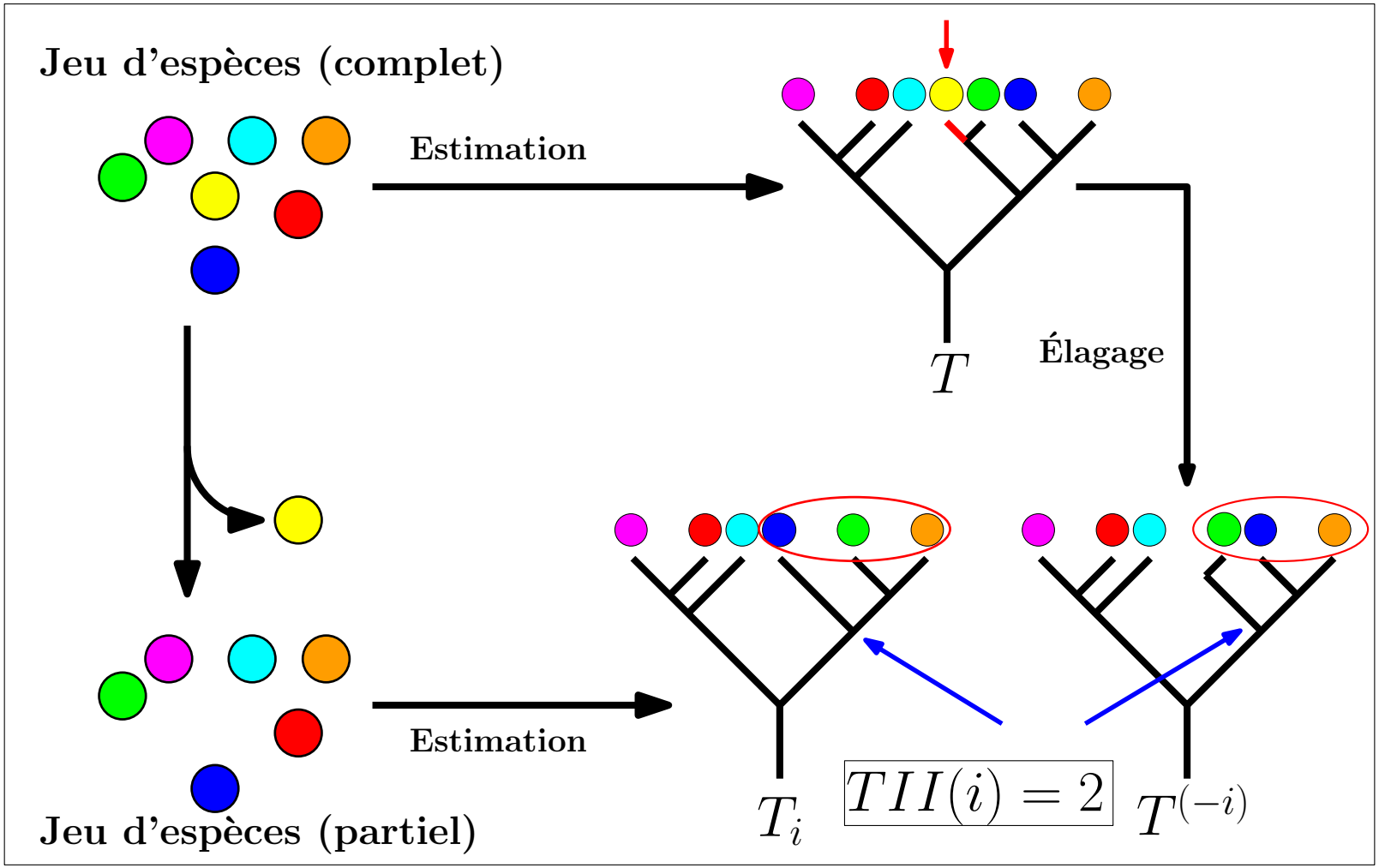
Taxon Influence Index (TII)

- **Objectif :** Étudier la stabilité de l'arbre vis-à-vis des espèces,
- **Justification :** Certaines espèces présentent des caractéristiques évolutives, non prises en compte par le modèle, qui peuvent biaiser la phylogénie estimée.
- **Méthode :**
  - Estimer la phylogénie T sur jeu complet d'espèces,
  - Retirer les espèces une par une et estimer le nouvel arbre T_i sur le nouveau jeu d'espèces,
  - Évaluer la différence entre T et T_i par $d(T, T_i)$.

Méthode


Jeu d'espèces (complet)
Estimation
T
Élagage
Estimation
Jeu d'espèces (partiel)
T_i
$T^{(-i)}$

Méthode


Jeu d'espèces (complet)
Estimation
T
Élagage
Estimation
Jeu d'espèces (partiel)
T_i
$TII(i) = 2$
$T^{(-i)}$

Les TII en pratique

Interprétation :

TII :

- Valeur basse : ajouter/retirer l'espèce du jeu de données n'a (presque) pas d'impact sur l'estimation,
- Valeur élevée : espèce influente dont l'ajout/retrait au jeu de données modifie notablement la phylogénie estimée.

Stratégie pour obtenir des arbres stables

- Se concentrer sur les espèces influentes : celles avec de forts TII,
- Les trier par TII décroissant,
- Les retirer une à la fois jusqu'à tomber sur un arbre stable.

Les TII en pratique

Interprétation :

TII :

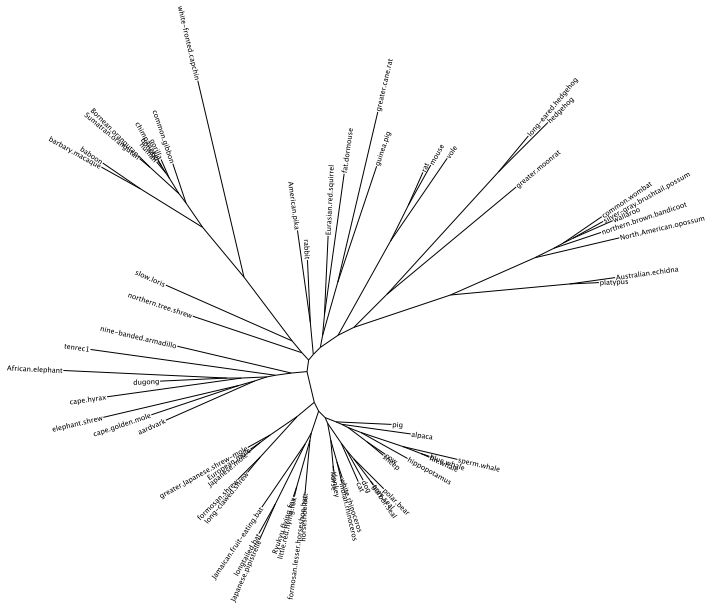
- Valeur basse : ajouter/retirer l'espèce du jeu de données n'a (presque) pas d'impact sur l'estimation,
- Valeur élevée : espèce influente dont l'ajout/retrait au jeu de données modifie notablement la phylogénie estimée.

Stratégie pour obtenir des arbres stables

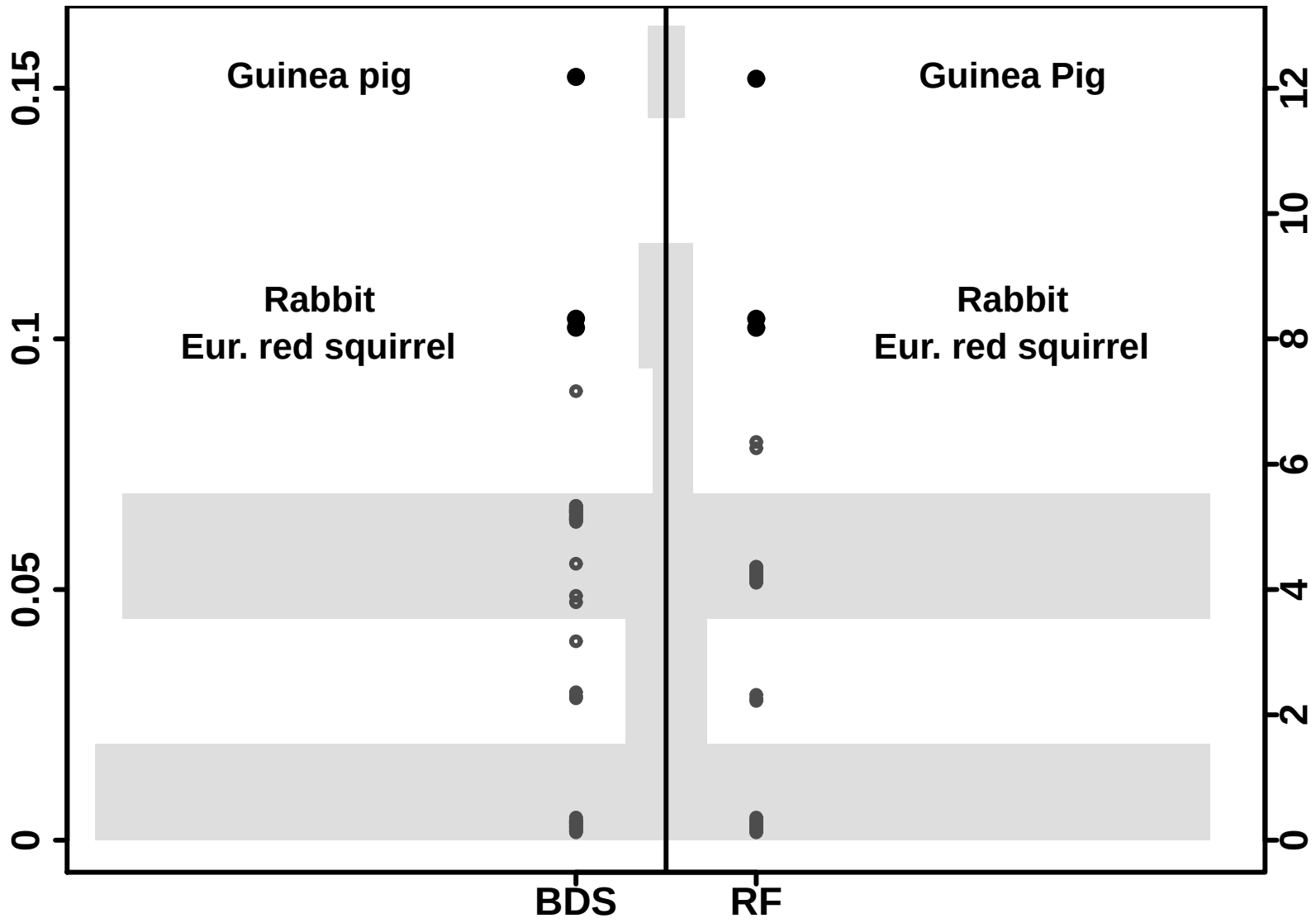
- Se concentrer sur les espèces influentes : celles avec de forts TII,
- Les trier par TII décroissant,
- Les retirer une à la fois jusqu'à tomber sur un arbre stable.

Données : mammifères à placenta

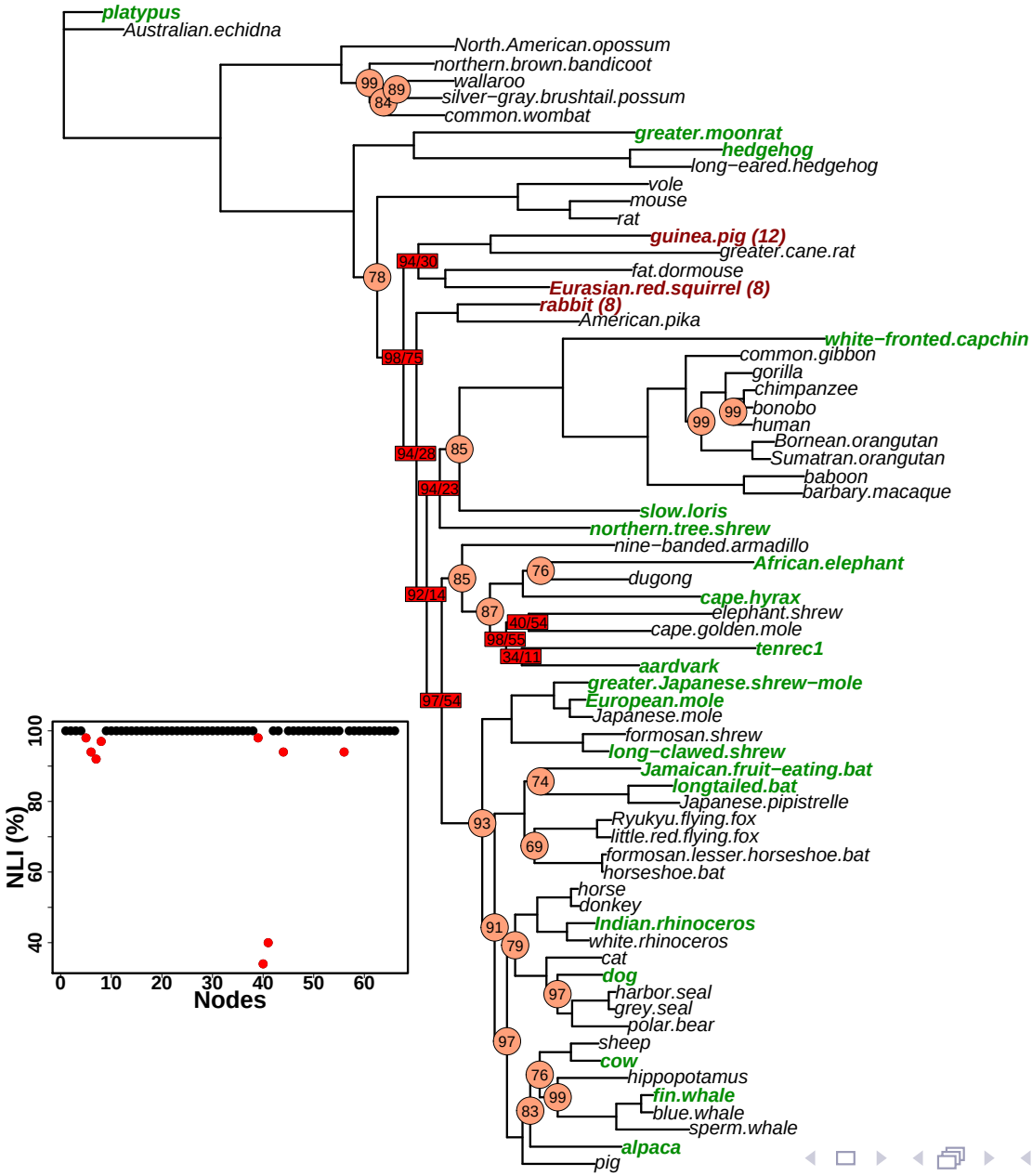
- Génome mitochondrial de 68 mammifères,
- Séquences d'acide aminés,
- Séquences de 3658 nucléotides,
- Phylogénie publiée par Nikaido *et al.* en 2003.



Taxon Influence Index



Phylogénie complète



Controverse sur le cochon d'inde

- Graur *et al.* : Is the guinea-pig a Rodent ? (1991)
- Hasegawa *et al.* : Rodent polyphily ? (1992)
- Cao *et al.* : Phylogenetic place of guinea-pigs : ... (1994)
- D'erchia *et al.* : The guinea-pig is not a rodent (1996)
- Philippe : Rodent monophyly : pitfalls of molecular phylogeny (1997)
- ...

Perspectives

Sources de variabilité

- Échantillonnage de sites $\rightarrow$ bootstrap ;
- Échantillonnage d'espèces $\rightarrow$ TII.

En cours

- Impact de la méthode d'inférence choisie ;
- Traitement des espèces influentes ;
- Seuil pour décréter qu'une espèce est influente.

Perspectives

Sources de variabilité

- Échantillonnage de sites $\rightarrow$ bootstrap ;
- Échantillonnage d'espèces $\rightarrow$ TII.

En cours

- Impact de la méthode d'inférence choisie ;
- Traitement des espèces influentes ;
- Seuil pour décréter qu'une espèce est influente.

Précision, répétabilité, robustesse

Un peu de vocabulaire

Précision : Probabilité que le noeud soit présent dans le vrai arbre ;

Répétabilité : Probabilité que le noeud soit retrouvé dans l'analyse de jeux de données indépendants.

Robustesse : Capacité à ne pas être affecté par de petites modifications dans les données ou les paramètres du modèle.

Bootstrap : mesure de répétabilité/précision ?

- Estimateur trop variable de la répétabilité (Hillis et Bull, 1993) ;
- Estimateur conservateur de la précision *id.* et (Susko, 2009) ;

Rogue Species